International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.8May - June – 2024
DOI:10.46647/ijetms.2024.v08i03.062 ISSN: 2581-4621

DETECTION OF LUNG CANCER FOR CLASSIFICATION APPROACH ON MINING TECHNIQUE

Gokul.K[1], Dr.R.Sankarasubramanian[2]
[1]*Research Scholar, Department of computer science, Erode Arts and science college, Erode-09.*
[2]*Principal, Erode Arts and science college, Erode-09.*

ABSTRACT
The leading cause of death for both men and women is cancer. A full recovery from cancer may be possible with early detection. Thus, there is a growing need for methods to identify carcinogenic nodules at an early stage. Lung cancer is one illness that is frequently misdiagnosed. Early detection of lung cancer saves countless lives; if treatment is delayed, more serious issues could arise that could result in an abrupt death. Numerous applications of knowledge discovery and data mining can be explored in the scientific and business domains. The application of data mining techniques in the healthcare system can yield valuable knowledge. In this work, we take a quick look at how classification-based data mining methods like BFO, SVM, LDA, and neural networks might be used to enormous amounts of medical data. Large volumes of healthcare data are gathered by the healthcare sector, but sadly, they are not "mined" to find secret information. It can forecast a patient's risk of developing lung cancer using common symptoms like age, sex, wheezing, shortness of breath, and pain in the arm, chest, or shoulder. This study provides an overview of the ongoing research on lung cancer that is being conducted with data mining tools. The purpose of the paper is to provide a model for early disease detection and accurate diagnosis, which will enable the physician to preserve the patient's life.
Keywords: Data Mining, Classification, KDD, Data Mining methods.

1 INTRODUCTION

Data mining is the process of analyzing the patterns of data that are hidden natural, according to different perspective for categorization of relevant information, which is collected and assembled in data warehouses, for effective analysis. Mining algorithm facilitates decision making and other information requirement to unlimitedly reduce the costs and increase the turnover. It is also known as ‘data in knowledge discovery’. Apart from analysis, it involves process such as data pre-processing of discovered patterns visualization and online updating. The other terms that resemble the mining are ‘Data dredging, Data fishing and Data snooping’ Stages in Mining

1.1 DATA MINING

Machine learning (ML) has several uses, but data mining is the most important one. One powerful new technique that has developed and grown so quickly is data mining, or the analysis step of knowledge discovery in databases. There is a lot of promise in this technology for businesses to help them concentrate on the most crucial details of the data they need to gather in order to understand the behavior of their customers. By going through several steps, such as "data selection, cleaning, data integration, transformation, and pattern extraction," intelligent approaches are used to extract patterns from the data. Classification, Regression, Clustering, Rule generation, Discovering, Association Rule, and other techniques are used to extract data, and each has a unique set of algorithms that it uses to try to fit into a data model**.** The field of data mining developed as a means of extracting information and knowledge from databases to discover patterns or concepts that are not evident.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.8May - June – 2024
DOI:10.46647/ijetms.2024.v08i03.062 ISSN: 2581-4621

2. KDD

Data mining (DM), also known as ‘‘knowledge discovery in databases” (KDD), is the process of discovering meaningful patterns in huge databases .The terms Knowledge Discovery in Databases (KDD) and Data Mining are often used interchangeably. KDD is the process of turning the low-level data into high-level knowledge. Hence, KDD refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. Although data mining and KDD are sometimes used interchangeably, data mining is actually a crucial stage in the KDD process. Data mining is depicted as a step in an iterative process of knowledge discovery in the following figure 1.

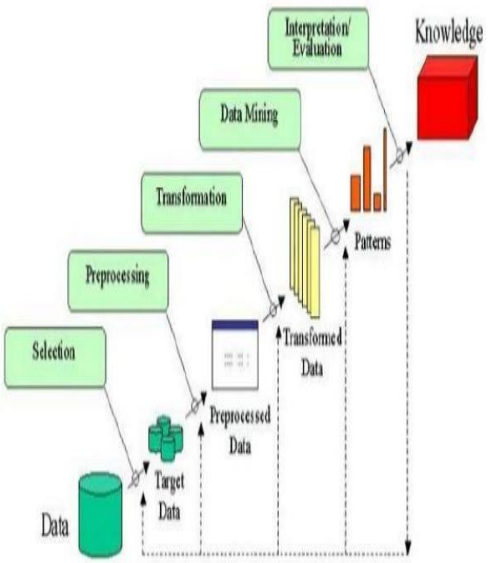

Figure 1 The KDD Process

- **Stores data**: Where data are stored in any data repository.
- **Data integration**: Multiple or heterogeneous data sources are integrated into a single unit.
- **Data selection**: Data retrieved from the database as needed for the KDD method.
- **Data transformation**: It is a process of data normalization where data are transformed and joined together into a form that is appropriate for the mining process. Sub stages of this data transformation are,
- **Data cleaning**: It handles noisy, erroneous and irrelevant data.
- **Improve data**: Improve quality of data by adding new information, missing values to available data.
- **Generalizing data**: Applying operations on data in order to prepare for a machine learning approach.
- **Data mining**: This is a very important step in the mining process. Here intelligent methods are applied in order to extract data patterns.
- **Pattern evaluation**: This process is to identify the truly interesting patterns that are presented in the knowledgebase.
- **Knowledge presentation**: It is the final reporting of the KDD process. Where visualization and knowledge representation techniques are used to represent the mined knowledge to the user.

3. DATA MINING CLASSIFICATION METHODS

Numerous techniques are used in data mining. Every method has its own set of benefits and drawbacks and serves a distinct function. Since the applicable prediction algorithms assign patients to either a "benign" group, which is non-cancerous, or a "malignant" group, which is cancerous, and develop rules for the same, the majority of data mining methods widely employed for this review

DOI:10.46647/ijetms.2024.v08i03.062 ISSN: 2581-4621

fall into this categorization category. Accordingly, the diagnostic issues related to lung cancer fall mostly under the category of the much debated classification issues. One of the most crucial tasks in data mining is classification. The data is mapped to predetermined goals. Because the goals are predetermined, it is guided learning. Building a classifier based on some examples with various qualities to describe the objects or one attribute to define the group of objects is the classification's goal. Based on the values of other attributes, the classifier is then used to predict the group attributes of new examples from the domain. The following categories apply to the widely used techniques for data mining classification tasks [4].

3.1 SUPPORT VECTOR MACHINE (SVM)

Support vector machine (SVM) is an algorithm that attempts to find a linear separator (hyper-plane) between the data points of two classes in multidimensional space. SVMs are well suited to dealing with interactions among features and redundant features.

3.2. BACTERIAL FORAGING OPTIMIZATION

Passion has presented an optimization technique called bacterial foraging optimization algorithm, which takes its cue from the way bacteria like E. Coli forage for food. If E. Coli is given the right conditions and enough food to flourish, it grows very quickly. The E. Coli bacteria travels quickly toward places rich in nutrients and attempts to avoid unpleasant materials. Bacteria are known to move in what are called taxes. According to foraging theory, the goal is to find and gather nutrients in a way that minimizes energy intake per unit of time (E/T). An E. Coli bacterium moves its 8–10 flagella at variable speeds of 100–200 rps throughout its body. The spinning flagella, which serve as a biological motor and aid in the swimming of bacteria, define its movement and behavior. Four consecutive mechanisms make up the majority of the bacterial foraging process: chemo taxis, swarming, reproduction, and elimination dispersal.

3.3 NEURAL NETWORKS

Systems that are designed after how the human brain functions are known as neural networks (NN). Similar to how synapses connect the millions of neurons that make up the human brain, a neural network is a collection of connected input/output units where each connection has a weight. During the learning phase, the network learns by modifying the weights in order to predict the input's right class label.

4. DATA MINING CLASSIFICATION TECHNIQUES FOR LUNG CANCER DETECTION

The study conducted by Yongqian Qiang, YouminGuo, Xue Li, Qiuping Wang, Hao Chen, and Duwu Cuic [1] aimed to investigate novel approaches in the diagnosis of peripheral lung cancer through data mining techniques. Additionally, the study sought to acquire early-stage computer-aided detection (CAD) technology and knowledge support. Once the properties of the clinical and CT findings were identified, the data were loaded into the database. The three data mining technologies' diagnosis guidelines for peripheral lung cancer are identical to clinical diagnostic guidelines, and they may also be utilized to expand the expert system's knowledge base. Data mining technology has been shown to have potential benefits in differential diagnosis and clinical imaging diagnosis.

Cheng-Mei Chen and Chien-Yeh Hsu [2] established a survival prediction model for liver cancer using data mining technology. They collected the data from the cancer patient’s registration database of a medical

centre in Northern Taiwan between the years 2004 and 2008. A total of 227 patients were newly diagnosed with liver cancer during this time. They extracted nine variables pertaining to liver cancer survival were analysed using t test and chi-square test through literature survey and expert consultation. Six variables showed significant. Artificial neural networks (ANN) and classification and regression trees (CART) were adopted as prediction models.

DOI:10.46647/ijetms.2024.v08i03.062 ISSN: 2581-4621

Based on fuzzy rough set theory, Dai and Xu [3] suggested a dimension reduction approach with respect to fuzzy gain ratio. Three gene expression data sets for actual tumors were used. The effectiveness of their model was demonstrated in the paper along with its classification accuracy.

According to research by Thangaraju P, Karthikeyan T, and Barkavi G [4], smoking is the main cause of lung cancer risk. The chance of acquiring lung cancer increases with the number of years and cigarettes smoked. Although those who are younger can develop lung cancer, the average age of a lung cancer diagnosis is 65 to 70 years old. Lung cancer can also strike young persons who have never smoked.

In order to predict the presence or absence of lung cancer, identify marker genes linked to the particular type of the disease, and predict the specific type of lung cancer, Jennifer et al. [5] proposed a system that uses gene expression data from nucleotide microarrays. The suggested method would function as a trustworthy supplement to the current techniques for classifying lung cancer, aiding in a quicker detection.

According to past studies and case studies, cancer patients have a greater chance of surviving when the disease is detected early on, as stated by Julliet et al. [6]. Researchers are pursuing data mining approaches for pre-diagnosis because lung cancer is a disease that heavily depends on previous data for early diagnosis. With early identification at stage 1, when the tumor has not yet spread, the five-year survival rate rises to 70%. In addition to being expensive and requiring complicated equipment, current medical procedures including X-rays, CT scans, sputum cytology analyses, and other imaging techniques have only been shown to be effective when the tumour has spread to other body areas, or stage 4.

According to Wenyan et al. [7], the Fisher discriminate model, Random Forests algorithm (RF), and Support Vector Machine (SVM) are effective supplemental diagnosis techniques for lung cancer. Given that the SVM and RF algorithms' diagnosis indices outperform Fisher discriminate analysis, it is plausible that they are determining the best lung cancer classification model. The study on lung cancer diagnosis using SERS on data mining can be a new form of lung cancer diagnosis tool, according to the results when compared to healthy individuals.

Iyrena et.al [8], investigates application of novel Bidirectional Data Partitioning Technique (BDP) to cancer survival analysis. Author has developed this technique for classification problems with unstable feature relevance and SEER Cancer Data illustrates this machine learning concept. BDP is applied for survival analysis in order to find groups of patients with different key factors that determine survival time.

5. METHODOLOGY

First, various techniques for categorization are selected from data mining and put into practice using a computer language. Here, MATLAB will be used as the implementation language. Data about lung cancer will be used as input data for each algorithm's test. I am utilizing many classification algorithms in addition to optimization techniques.

6. CONCLUSION

The presented discussion on knowledge extraction from medical databases is merely a short summary of the ongoing efforts in this area. However, point to interesting directions of our research, where the aim is to apply hybrid classification schemes and create data mining tools well suited to the crucial demands of medical diagnostic systems. It is proposed to develop a substantial set of techniques for computational treatment of these data. The approaches in review are diverse in data mining methods and user interfaces and also demonstrate that the field and its tools are ready to be fully exploited in biomedical research.

DOI:10.46647/ijetms.2024.v08i03.062 ISSN: 2581-4621

REFERENCES

[1] YonqianQia,ng, YouminGuo, Xue Li, Qiuping Wang, Hao Chen, & DuwuCuic, "The Diagnostic Rules of Peripheral Lung cancer Preliminary study based on Data Mining Technique", Journal of Nanjing Medical University, Vol. 21(3), pp. 190-195.

[2] Cheng-Mei Chen, Chien-Yeh Hsu, Cheng-Mei Chen andChien-Yeh Hsu," Prediction of Survival in Patients with Liver Cancer using Artificial Neural Networks and Classification and Regression Trees", Seventh International Conference on Natural Computation 2011.

[3] Jianhua Dai, Qing Xu," Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumour classification", Applied Soft Computing,Vol.13(2013), pp. 211–221.

[4] Thangaraju P, Karthikeyan T, BarkaviG,"Mining Lung Cancer Data for Smokers and Non-Smokers by Using Data Mining Techniques", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 7, July 2014.

[5] J. Cabrera, A. Dionisio and G. Solano, "Lung cancer classification tool using microarray data and support vector machines," Information, Intelligence, Systems and Applications (IISA), 2015 6th International Conference on, Corfu, 2015, pp. 1-6.