# Decision Support System for Dyspnoea Related Diseases Diagnosis

Nan Yin Wai Toe Myint, Mu Mu Myint
*University of Computer Studies, Yangon*
*nanyinwai@gmail.com, nanyinwai.ucsy@gmail.com*

## Abstract

*In medicine, diagnosis is the process of identifying a medical condition or disease by its signs and symptoms. The decision support system (DSS) is used to refer computer systems that offer information to the junior doctors in a more flexible form. This system use Bayesian classification to find the all probability of all diseases and to generate the maximum probability of disease which the patient suffers. Dyspnoea related diseases include (25) attributes (symptoms) and (18) classes (diseases). In this paper, a decision support system is implemented by using Bayesian Analysis to search the similar pattern of disease. A typical decision support system is designed to assist the junior doctors to make decisions.*

## 1. Introduction

The main purpose of clinical decision support system (CDSS) is to assist junior doctors. This means that a junior doctor would interact with a CDSS to determine diagnosis, analysis for patient data. The junior doctor would input the information and wait for the CDSS to output the “right” choice and the junior doctor would simply act on that output. Typically the CDSS would make suggestions of outputs or a set of outputs for the junior doctor to look through and the junior doctor officially picks useful information and removes erroneous CDSS suggestions.

Dyspnoea is the subjective sensation of shortness of breath, often exacerbated by exertion. It is a common symptom of cardiac and respiratory diseases, but it may occur as a result of other diseases, e.g. - severe anemia, acidosis, and may be psychogenic. It is often difficult to differentiate dyspnoea due to heart disease from that caused by lung disease. A history of cough, wheezing and nocturnal dyspnoea may occur in cardiac failure as well as in patient with lung disease. Dyspnoea can vary from dyspnoea on exertion to dyspnoea at rest and even patient can’t speak.

Dyspnoea arises from cardiac disease has different characters such as orthopnoea and paroxysmal nocturnal. Most of dyspnoea is aggravated by exertion, stress and some medicine causing salt and water retention). Acute severe dyspnoea is one of the most common medical emergencies. Although there are usually a number of possible causes, attention to the history. Careful examination suggests dyspnoea related diseases diagnosis which can be confirmed by routine investigations.

In this system, all the symptoms of dyspnoea patients' records and their diseases are used as train datasets (old history records). If new patient with dyspnoea arrives, the system asks the pattern of dyspnoea related diseases and diagnoses the related diseases.

The rest of this paper is organized as follows. Section 2 describes related work. Section 3 describes proposed system. Section 4 describes experimental result. Finally, concludes this paper in section 5.

## 2. Related Work

There are many advantages use of machine learning, such as training a medical diagnosis system on data from all hospitals in the world [6]. If diagnosis functions are different in medical cases, methods such as hierarchical Bayesian approaches provide one way to tackle this problem.

The development classification methods are explained by [11]**.** In the classification method, training set is used by the classification programs to learn how to classify objects. There are two phases for constructing a classifier, training set and testing set. The training set is used to decide how the parameters ought to be weighted and combined in order to separate the various classes of objects. The testing set is used to get the classification accuracy for diseases.

The use of clinical decision support system in healthcare is presented by [9]**.** Software that integrates information on the characteristics of patients with a computerized knowledge base for the purpose of generating patient-specific assessments or recommendations designed to aid junior doctors or patients in making clinical decisions.

Machine learning techniques are used in medicine for diagnosis whether a patient suffers from a particular disease or not using records for already observed patients with known diagnosis [8].

## 3. Proposed System

Medical diagnosis is a complex human process that is difficult to represent in an algorithmic model. Medical diagnosing requires the understanding of symptoms, drug-drug interactions and patient history. The diagnosing process requires knowledge of diseases in general. The development of an effective clinical decision support system will have a significant impact on practice methodology. Clinical decision support systems are intended to receive patient data and utilize that data to propose a series of possible diagnoses and a course of action. Without CDSS may lead to the misdiagnosis of the patient. In this paper, the use of clinical decision support system will help junior doctors in order to get quick and accurate diagnosis for dyspnoea related diseases.

This paper includes the developed clinical decision support system based on data mining aspect. The most common data mining tasks are estimation, prediction, description, clustering, association and classification. Among of many data mining tasks, Classification methods from statistical pattern recognition, neural nets, and machine learning were applied to real-world data sets. Each of these data sets has been previously analyzed and reported in the statistical, medical, or machine learning literature. The data sets are characterized by statically uncertainty; there is no completely accurate solution to these problems. Training and testing or resembling techniques are used to estimate the true error rates of the classification methods.

There are (25) attributes (symptoms) which can identify (18) classes (diseases). The following diseases are used in these systems which are common presentation of dyspnoea related diseases.

- Pneumonia
- Acute Viral Infection
- Acute Pulmonary Oedema
- Chronic Obstructive Pulmonary Disease (COPD)
- Asthma
- Inhaled Foreign Body
- Pulmonary Embolism
- Acute Myocardial Ischaemia
- Pneumothorax
- Diabetic Ketoacidosis
- Psychogenic
- Milliary TB
- Large Pleural Effusion
- Diabetic Mellitus ( DM )
- Obesity
- Severe Anaemia
- Lungs Cancer ( CA Lungs)
- Chronic Heart Failure

Figure 1 shows the system flow of the proposed system by using Naïve Bayesian Classifier.

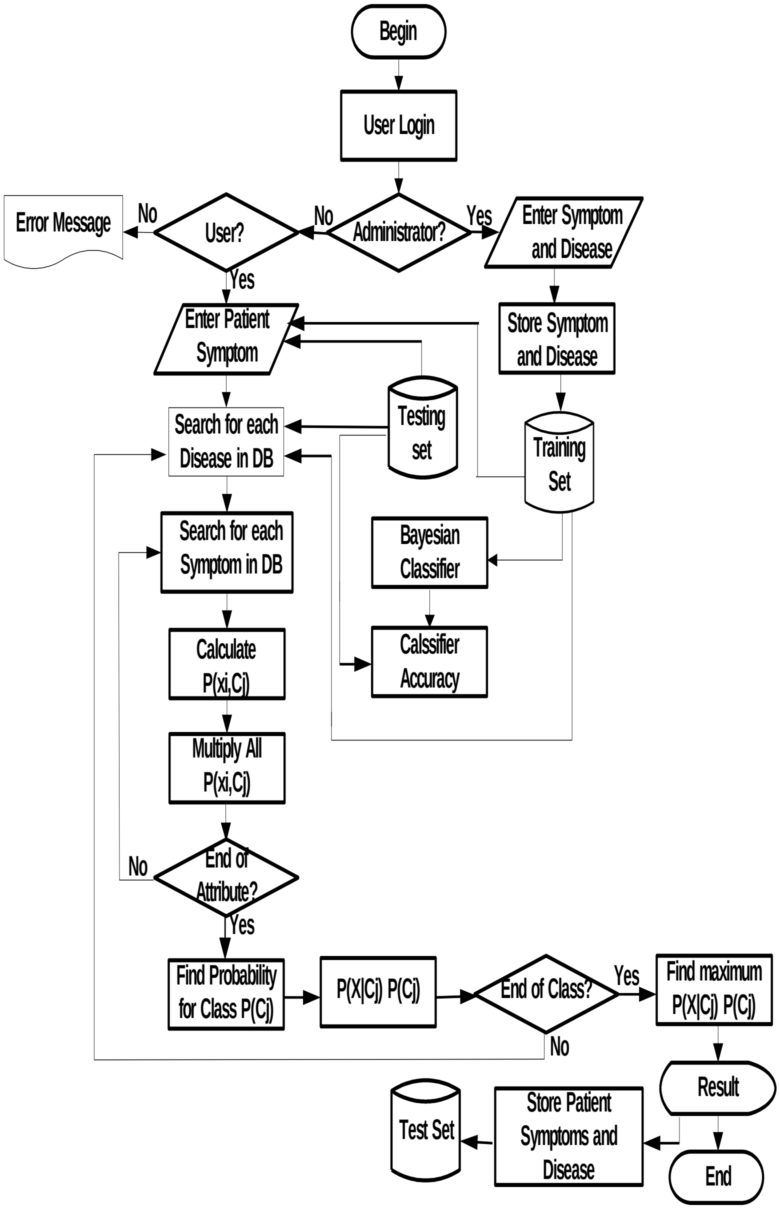

**Figure1.System Flow of the Proposed System**

The doctor will ask the patient who suffered dyspnoea related diseases based on the above twenty-five symptoms. For example, one of the patients who suffered dyspnoea comes to the hospital. He suffer cough in early morning. He suffer aggravating when he get infection, and onset (duration) is acute. His sputum is mucoid and he relieve when he takes drugs. He is a smoker and his occupation has stress. He suffer also wheezing when he sleep**.** Sometime he get allergy when he eat something or drink alcohol and suffer cyanosis symptom**.** There are two people who suffer dyspnoea in his family history. This system matched and calculated the patient symptoms with the above 1003 old history records and my get the result with the maximum probability of disease. So, the result may be Asthma disease for that person who came with above symptoms.

In this paper, Onset means duration of dyspnoea. It can be divided into two types such as Acute and Chronic. Their attribute value is Yes or No. Fever can be divided into High Grade Fever and Low Grade Fever. Their attribute value is Yes or No. Cough can be divided into Anytime, Nocturnal (night) and Early Morning. Character of Chest Pain can be distinguished into Central, Peripheral, and Central and Peripheral. Their attribute value is Yes or No. Aggravating Factor means that how patient can aggravate from dyspnoea related diseases. That can be characterized into Exertion, Infection, Drugs and Smoking. Their attribute value is Yes or No. Relieving Factor means that how patient can relieve from dyspnoea related diseases. That can be characterized into Oxygen, Drugs and Rest. Their attribute value is Yes or No. Character of Sputum can be distinguished into Mucopurulent or Rusty, Serous and Mucoid. Associated Factor consist of Even At Rest, Haemoptysis, Wheezing, Stridor, Occupational, Personal Problem, Smoking, Orthopnoea, Paroxysmal Nocturnal Dyspnoea, Palpitation, Oedema, Sputum, Calf Muscle Pain, Allergy, Cyanosis, Family History, Vomiting, Abdominal Pain and Jaundice. Their attribute value is also Yes or No.

## 3.1. Bayesian Classification

Bayesian classification is supervised learning method. It is a statistical method for classification. This method can be assumed as an underlying probabilistic model. Bayesian classification allows us to capture uncertainty about the model in a principle way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems.

## 3.2. Process Flow of the Naïve Bayesian Classifier

The processes of this system using Naïve Bayesian analysis is as follows:

**(1)** User symptoms are applied to the system through the user interface.

**(2)** Those symptoms are changed to attribute vector in order to apply to Naïve Bayesian Classifier to classify the disease.

$X = x_1, x_2, x_3, \ldots, x_n$, where n is the number of symptoms (attributes).

**(3)** Then the classifier starts working. It scans all possible diseases from the training data sets. Those diseases will be the labeled classes.

$C = C_1, C_2, \ldots, C_m$, where m is the number of diseases (classes).

**(4)** Find the probability for each Class

$P(C_j)$, where j=1 to m, for each attribute in the user's symptoms,

**(5)** Find the Probability of Symptom-based Class Label.

$P(x_i \mid C_j)$, where i = 1 .. n, number of attributes in the input data set (user's symptoms), j = 1 .. m, where number of class labels in the training set.

**(6)** $P(X \mid C_j)$ is calculated for all attributes in the input data set (testing set).

$$P(X \mid C_j) = \Pi \; P(x_i \mid C_j)$$

**(7)** In order to classify the unknown input data set X, $P(X \mid C_j)\,P(C_j)$ is evaluated for each class $C_j$. Sample X is then assigned to the class $C_j$ if and only if

$P(X \mid C_j)\,P(C_j) > P(X \mid C_i)\,P(C_i)$ for 1 <= i <= m, i <> j.

**(8)** The class of the input set would be the one with maximum probability from above step [10].

## 3.3. Naïve Bayesian Analysis

$$P\ (d|s) = \frac{P(d)*P(s|d)}{P(s)}$$

The probability of a disease given a symptom P (d|s) is dependent on the probability of that anyone in the population has the disease P(d), has the symptom P(s) and the likelihood that given the disease the probability of having the symptom is P(s|d).

## 4. Experimental Results

In the dyspnoea related diseases diagnosis system, we diagnose eighteen diseases based on twenty-five symptoms according to 1003 real records. And then spilt the original patient dataset into a training set and a test set and keep percentages of the positive and negative samples same in the training and test sets. The training set is used to derive the classifier, whose accuracy is estimated with the test set. We summarize some basic information about the datasets, including the number of features, the sizes of the training and test sets.

The accuracy is generated using Holdout Method. In the holdout method, the given data are partitioned into two independent sets, called the training set and the testing set. Two thirds of the data are allocated to the training set, and remaining one third is allocated to the test set. The training set is used to derive the classifier, whose accuracy is estimated with the test set shown in Table 1.

**Table 1. Sample Training Dataset**

| Onset | Chest Pain | Wheezing | Cough | Aggravating Factor | Relieving Factor | Sputum | Diseases |
|---|---|---|---|---|---|---|---|
| Acute | No | Yes | Early Morning | Infection | Drugs | Mucoid | Asthma |
| Acute | Central | No | Night | Exertion | Drugs | Serous | Acute Pulmonary Oedema |
| Chronic | No | No | Anytime | No | No | No | Large Pleural Effusion |
| Chronic | No | Yes | Anytime | Drugs | Rest | No | Chronic Heart Failure |
| Acute | No | Yes | Night | Exertion | Drugs | Serous | Acute Pulmonary Oedema |

Calculation of the system is shown as following. New Patient comes with the Symptoms as below:

X=Onset(x1) =”Acute”, Chest Pain (x2) =“No”, Wheezing(x3) =“Yes”. What will be the Disease?

Step1: Calculate the probability of each disease in record.
P (Asthma) = 1/5 = 0.2
P (Acute Pulmonary Oedema) = 2/5 = 0.4
P (Large Pleural Effusion) = 1/5 = 0.2
P (Chronic Heart Failure) = 1/5 = 0.2

Step2: Calculate the probability of each symptom for disease in record.
P (Onset=“Acute”) = 1/1 = 1……Asthma
P (Onset=“Acute”) = 2/2 = 1……Acute Pulmonary Oedema
P (Onset=“Acute”) = 0/1 = 0……Large Pleural Effusion
P (Onset=“Acute”) = 0/1 = 0……Chronic Heart Failure

P (Chest Pain=“No”) = 1/1 = 1...….Asthma
P (Chest Pain=“No”) = 1/2 = 0.5….Acute Pulmonary Oedema
P (Chest Pain=“No”) = 1/1 = 1 ……Large Pleural Effusion
P (Chest Pain=“No”) = 1/1 = 1 ……Chronic Heart Failure
P (Wheezing =“Yes”) = 1/1 = 1..…..Asthma
P (Wheezing = “Yes”) = 1/2 = 0.5.....Acute Pulmonary Oedema
P (Wheezing = “Yes”) = 0/1 = 0 …...Large Pleural Effusion
P (Wheezing = “Yes”) = 1/1 = 1 …...Chronic Heart Failure

Step 3: Multiply for each disease
Asthma = 1 * 1 * 1 = 1
Acute Pulmonary Oedema = 1 * 0.5 * 0.5 = 0.25
Large Pleural Effusion = 0 * 1 * 0 = 0
Chronic Heart Failure = 0 * 1 * 1 = 0

Step 4: Multiply with Prior Probability from step1
Asthma = 0.2 * 1 = 0.2
Acute Pulmonary Oedema = 0.4 * 0.25 = 0.1
Large Pleural Effusion = 0.2 * 0 = 0
Chronic Heart Failure = 0.2 * 0 = 0

Step5. The maximum Probability is 0. 2. So, the result is Asthma.

The accuracy measurement of the disease, Asthma is 99.0228013029316% which is tested upon 35 patients using 1003 all datasets.

**Table 2**

| Training Data | Classification Accuracy |
|---|---|
| 300 | 99.0228013029316% |
| 500 | 99.0228013029316% |
| 1003 | 99.0228013029316% |

When data are entered in the training data set the classification accuracy is not changed. Because new data are the same as from attribute values in the training data set.

**Table 3**

| Testing Data | Classification Accuracy |
|---|---|
| 306 | 99.0228013029316% |
| 250 | 98.125% |
| 300 | 95.3125% |

The classification accuracy is changed when data are entered in the testing data sets. Because data are different from attributes values in the training data set.

## 5. Conclusion

Solving the dyspnoea related diseases diagnosis is complex task. In medical diagnosis and prediction the use of clinical decision support system will make the junior doctors to be more interactive and save cost and time. Bayesian classifier is used to overcome the similarity problem which is a great handler in accurate and quick diagnosis of the dyspnoea related diseases. The aim of the paper is to prevent delay and the possibility of wrong treatment for patient before proper laboratory tests are conducted and their results are received.

## References

[1] http://en.wikipedia.org/wiki/clinical_decision_support_system

[2] Enrico Coiera, "Guide to Health Informatics 2nd Edition", 2003.

[3] Fahhad Farukhi, "Clinical Decision Support Systems", Public Health. Management & Policy, May 16, 2000.

[4] Sholom M. Weiss and Ioannis Kapouleas, "An Empirical Comparison of Pattern Recognition, Neural Nets, and Machine Learning Classification Methods", Department of Computer Science, Rutgers University, New Brunswick, NJ 08903.

[5] www.wrongdiagnosis.com

[6] Tom M. Mitchell, "The Discipline of Machine Learning", July 2006

[7] Chethan J S, Gayatri Ravichandran Geeta ,Joshwini Pereira, Krupa Jakkula, "Data Mining Concepts and Techniques" , 2007

[8] Zekie Shevked, Ludmil Dakovski, PhD student, Technical University Sofia, "Learning and classification with prime implicants applied to medical data diagnosis", 2007.

[9] Dawn Weathersby, MSN, RN, Quality Improvement Advisor, Informatics Nurse Specialist, "Healthcare decision support systems"

[10] Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques".

[11] Thin Ei Phyo, "Decision Support System for Acute Abdominal Pain", University of Computer Studies, Yangon, August 2007.