การจำแนกประเภทอาการสำคัญในโรคหูโดยใช้เทคนิคเหมืองข้อมูล

นายนรินทร์ วัฒนสุสิน

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตร์มหาบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์และสารสนเทศ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2554

CLASSIFYING CHIEF COMPLAINT IN EAR DISEASES USING DATA MINING TECHNIQUES

Mr. Narin Watanasusin

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computer Science and Information
Department of Mathematics and Computer Science
Faculty of Science
Chulalongkorn University
Academic Year 2011

Thesis Title	CLASSIFYING CHIEF COMPLAINT IN EAR DISEASES USING DATA MINING TECHNIQUES
By	Mr. Narin Watanasusin
Field of Study	Computer Science and Information
Thesis Advisor	Siripun Sanguansintukul, Ph.D.

Accepted by the Faculty of Science, Chulalongkorn University in Partial Fulfillment of the Requirements for the Master's Degree

………………………………………….. Dean of the Faculty of Science
(Professor Supot Hannongbua, Dr.rer.nat)

THESIS COMMITTEE

………………………………………….. Chairman
(Professor Chidchanok Lursinsap, Ph.D)

………………………………………….. Thesis Advisor
(Siripun Sanguansintukul, Ph.D)

………………………………………….. External Examiner
(Assistant Professor Worasit Choochaiwattana, Ph.D)

นรินทร์ วัฒนสุสิน : การจำแนกประเภทอาการสำคัญในโรคหูโดยใช้เทคนิคเหมืองข้อมูล. (CLASSIFYING CHIEF COMPLAINT IN EAR DISEASES USING DATA MINING) อ. ที่ปรึกษาวิทยานิพนธ์หลัก: อาจารย์ ดร. สิริพันธุ์ สงวนสินธุกุล, 70 หน้า.

หูเป็นอวัยวะที่สำคัญในระบบการได้ยิน ซึ่งระบบการได้ยินมีความสลับซับซ้อนเป็นอย่างมาก แพทย์ต้องใช้ความพยายามในการสรุปโรคให้ถูกต้อง จากสัญญาณ อาการ และผลการทดสอบ เพื่อกำหนดข้อสันนิษฐานของการวินิจฉัยโรค ก่อนการรักษา ผู้ป่วยส่วนใหญ่ในงานวิจัยนี้เป็นผู้ป่วยหนัก เพราะฉะนั้นแพทย์จะตัดสินใจให้การรักษาด้วยวิธีการผ่าตัดมากกว่าการรักษาคนไข้ด้วยการให้ยา ผลที่ได้จากการจัดกลุ่ม เป็นสิ่งที่วิกฤตอย่างมากสำหรับแพทย์ สำหรับใช้เป็นข้อมูลสนับสนุนการวินิจฉัยโรคของแพทย์ก่อนที่จะดำเนินการผ่าตัดผู้ป่วย งานวิจัยนี้พยายามใช้ความสามารถอันชาญฉลาดของเทคนิคเหมืองข้อมูลที่จะค้นหารูปแบบที่ถูกซ่อนอยู่ในกลุ่มชุดข้อมูล ในที่นี้เทคนิคโครงข่ายปัญญาประดิษฐ์ เทคนิคนาอีฟเบย์ และต้นไม้ตัดสินใจ ถูกนำมาใช้เป็นเทคนิคในการจัดกลุ่มผู้ป่วยด้วยอาการหลักของโรคหู ผลลัพธ์ค่าความถูกต้องจากการจัดกลุ่มโรคหูจะมีความถูกต้อง ที่ 100% ในทุกเทคนิค

ภาควิชา คณิตศาสตร์ ลายมือชื่อนิสิต.................................
สาขาวิชา วิทยาการคอมพิวเตอร์และสารสนเทศลายมือชื่ออ.ที่ปรึกษาวิทยานิพนธ์หลัก.......
ปีการศึกษา 2554

5173617823 : MAJOR COMPUTER SCIENCE AND INFORMATION

KEYWORDS: DATA MINING / CLASIFICATION TECHNIQUE / EAR DISEASES

NARIN WATANASUSIN : CLASSIFYING CHIEF COMPLAINT IN EAR DISEASES USING DATA MINING

THESIS ADVISOR: SIRIPUN SANGUANSINTUKUL, Ph.D., 70 pp.

Ears are the important organ for the hearing system. The system itself is very complicated. The clinicians attempt to determine the correct diagnosis using signs, symptoms and test results to formulate the hypothesis of the diagnosis before providing treatments. Most patients in this study have severe illness. Therefore, the clinicians decide to take the treatment by surgery rather than treating the patients with medicine. The result of the classification is very critical for the clinicians to support their diagnosis before giving the surgery to the patients. This study endeavors on using intelligent capability of data mining to discover hidden patterns in the data. Here, Artificial Neural Networks (ANN), Naïve Bayes and Decision Tree are utilized as techniques to classify patients with chief complaints in ear diseases. The results of classifying the ear diseases are very encouraging with the percentage accuracy of 100% for all techniques.

Department: Mathematics	Student’s Signature ....
Field of Study : Computer Science and Information	Advisor’s Signature ....
Academic Year : 2011	

ACKNOWLEDGEMENTS

I would like to really appreciated to my thesis advisor, Siripun Sanguansintukul, PhD. for her helpful guidance, inspiration, great support and solving many problems. Without her suggestion and motivation while doing an experiment, this thesis could not be finished on time.

I would like to say “thank you very much “to all my classmate, and most importantly, my mother, father and family for everything that they have supported me and “thank you” Mr. Montree Imsamai for program consulting. The last one “Special thank you” Miss Narumol Nuanchawee from Operating Room Ear Eye Nose Throat Department of Ramathibodi Hospital for data set and consulting.

CONTENTS

PAGE

LIST OF TABLES

Tables Page

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Motivation and Problem Description

There are various methods in data mining that can be applied to classify tasks. This thesis presents the experiments in classifying patients with ear diseases. The data set consists of 35 variables on 257 patients. The objective of this experiment is to determine the ear diseases in patients. There are 6 targets class of ear disease family: Acute Serous Otitis Media, Chronic Serous Otitis Media, Otosclerisis, Cholesteatoma, Conductive Hearing Loss, and Sensorinueral Hearing Loss. Three selected data mining approaches are Artificial Neural Network, Naïve Bayes, and Decision tree. The results of classification might be useful in predicting the patients' disease and providing the suggestions for effective patients' treatments. Applying an intelligent [1] tool to classify the chief complaints can support the decision maker to diagnose the patients' symptoms in an effective way. To ensure that the patients would receive the appropriate treatments and patients would not become deaf. The technique with intelligent approaches may bring the huge impact to medical sector.

In general, medical diagnosis is the way to diagnose the disease [2]. The diagnosis includes checks-up history, physical examination, look, touch and listen to, laboratory tests, and special lab tests. Knowledge, skill, and experiences are required in order to diagnose and analyze what are the possible causes of patients' symptom. After that, the treatment is planned for the patients according to the symptoms. The human body is compound of several complex systems, such as brain system, movement system, visual system, auditory system etc. The disease that happens in each body's part requires a knowledgeable specialist to diagnose and suggest the treatment plan.

The hearing system is an important system of human body because it is necessary for human's communication system. The hearing system consists of a complicated organ. The patients who have hearing system problem are risk to deaf. The symptoms that relates to this case are severe patients and must treat by surgery. The symptoms on ear diseases are varied and some of them seem to be similar to other diseases on a facial system, such as earache, headache, dizziness, fever, and roaring. The clinicians who deal with ear diseases are specialist who attempt to determine ear disease from special signs and symptoms. Doctors must use their expertise, knowledge and medical tools to diagnose patients.

Data mining is an innovation tool that analyzes the large data set by using mathematic formula such as statistic, probability or matrix. Data mining acts as a process for knowledge discovery in databases (KDD). The knowledge of human starts from learning and remembers the whole things in brain that we call the experience. Learning makes the correct or wrong decision which depends on analysis of each person. Sometimes, we have to do trial and error for solving the problem that we do not know the solution.

1.2 The Objective of Research

The objective of this research is to assist doctors in sorting out and diagnosing ear disease, and providing a tool for doctors to distinguish the symptom of ear disease from other similar symptoms of other diseases. As a result, the physicians can diagnose patients accurately and quickly. In addition, the treatment can be planned and information can be further transfer to specialist.

1.3 Scope of Work

The data employed in this experiment are collected from Ear Nose Throat (ENT) department at Ramathibodi Hospital, Thailand. The data are obtained from 257 patients with ear problems. Patients came from all over the country with different genders, ages, and occupations. All patients already had a general checkup, and had been thoroughly diagnosed ear problems by doctors with special medical tools. When the doctors are sure that the patients have ear problems. The doctors will plan for a surgical treatment, then, patients will be admitted to the inpatient department.

Three techniques utilized as classifiers are Artificial Neural Network, Naïve Bayes and Decision Tree. These approaches are chosen because they are widely used and applied successfully in many applications [3] [4] [5].

MATLAB and SPSS are used as the simulation software tools in this research. MATLAB [6], MATrix LABoratory, is a software that collects all equations related in mathematics, statistic, physics, matrix and others. MATLAB suites for all users who want to deploy the experiment by themselves. The user should have skill in C language to operate such a program. SPSS [7], Statistical Package for the Social Sciences is a statistic application. This software is very easy to use because there are many toolboxes and wizards for utilizing the user. The GUI from both applications is not much different but MATLAB has more details than SPSS since MATLAB is able to display all outputs from the user's coding.

This thesis is structured as follows:

Chapter 2 provides the literature review. Chapter 3 describes the methodology. Chapter 4 discusses the experimental results. Chapter 5 gives the conclusion, discussion and future works.

CHAPTER 2
LITERATURE REVIEW AND BACKGROUND

2.1 Literature Review

Data mining are widely accepted as the technique for analyzing a large database. It has been used in various applications as follows:

Fei Fei Wang, Siripun Sanguasintukul and Chidchanok Lursinsap [8] proposed the artificial neural network technique as the classifier to predict the paper‘s curl in a papermaking industry. The experiment is run by SAS software. The original data set consists of 328 attributes. Principal component analysis and Step-wise regression are employed as reduced dimension technique. There are 41 attributes obtained from principal component analysis technique and 23 attributes are collected from step-wise regression technique. The double dogleg algorithm and Quasi-Newton algorithm are provided as the learning algorithms. The results are then compared. The raw data set was divided into two groups: train set and test set. The ratio is 60:40. The artificial neural network with Quasi-Newton algorithm was selected as the final model for the forecasting process. The bagging technique was used as a measurement to improve the accuracy of the final result. Finally, the accuracy percentage is 96.35%.

Naroumon Yordphet and Siripun Sanguansintukul [9] proposed the artificial neural network as the classifier for predicting the safety stock in a jewelry industry. All experiments are run by WEKA software. The data set consists of 372 records. The data set is divided into two groups: train set and test set. The ratio is 70:30. Therefore, there are 260 records in the train set and 112 records in the test sets. The architecture of the network is 8-3-1: 8 for input nodes, 3 for hidden nodes and 1 for output node. The other two important parameters: learning rate and momentum are set to 0.1 and 0.2, consequently. The number of epochs is set to 50000. The accuracy result is 97.37%.

Yi Wang, Siripun Sanguansintukul and Chidchanok Lursinsap [10] proposed the artificial neural network as a tool to predict the customer life value (CLV) for the mobile business. The data set consists of 12005 records and 126 different attributes. The data set is divided into two sets: train set and test set. The ratio is 60:40. The principal component analysis (PCA) is used as the reduced dimension method. After PCA, the new input consists of 27 attributes. All experiment in this paper is run by SAS application. The accuracy from experiment is 96.5%.

Sirilak Areerachakul and Siripun Sanguansintukul [11] proposed to classify the water quality using the artificial neural network. From 288 canals in Bangkok, the raw data set was collected from year 2003 – 2007. There are 11,820 records. Each record consists of 3 chemical factors: pH value, Dissolved Oxygen and Biochemical, and Oxygen Demand as input attributes. Five different classes were used as the measurement of water quality. The ratio of the train set to the test set is 60:40. Therefore, there are 7,092 records were used as the train set and 4,728 records were used as the test set. The architecture is 3-4-5, 3 is the number of input nodes, 4 is the number of hidden nodes, and 5 is the number of output nodes. The accuracy is 99.34%.

J. Víctor Marcos, Roberto Hornero, Daniel Álvarez, Félix Del Campo and Miguel López [12] classified the Obstructive Sleep Apnea Syndrome patient (OSAS) with the artificial neural network. In this study, they compared two different neural network models: multi-layer perceptron (MLP) and radial basis function (RBF) networks. The data set consists of 187 records: 76 records for negative diagnosis results and 111 records for positive diagnosis results. Each record consists of 5 attributes and the last attribute is the attribute class. The data set was collected from the Hospital Clínico de Santiago de Compostela in Spain. In this study, the training set and the test set consists of 113 and 74 objects respectively. OSAS is a respiratory disorder characterized by recurring episodes of upper airway occlusion during sleep.

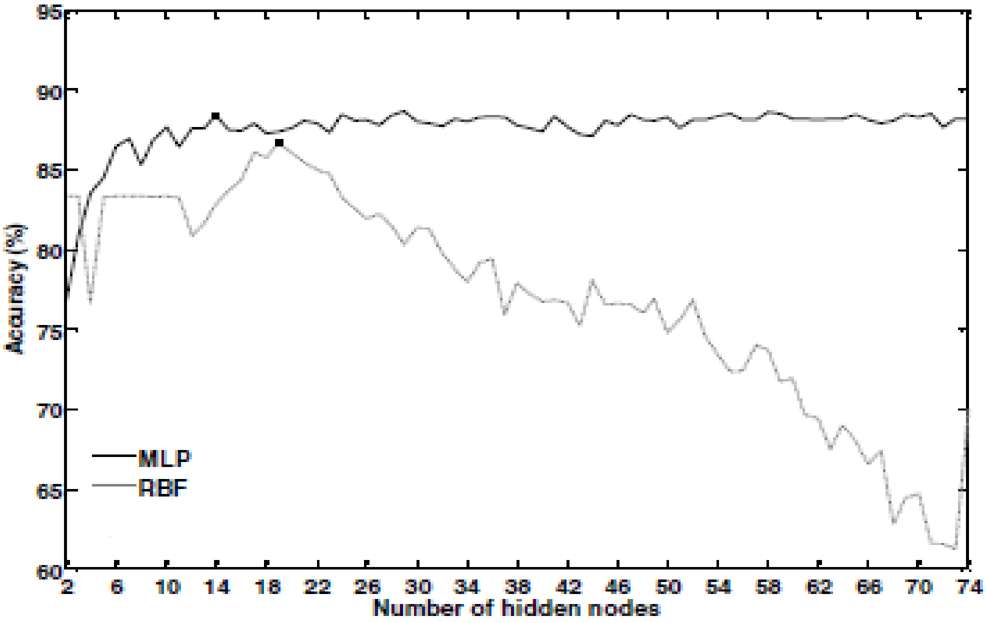
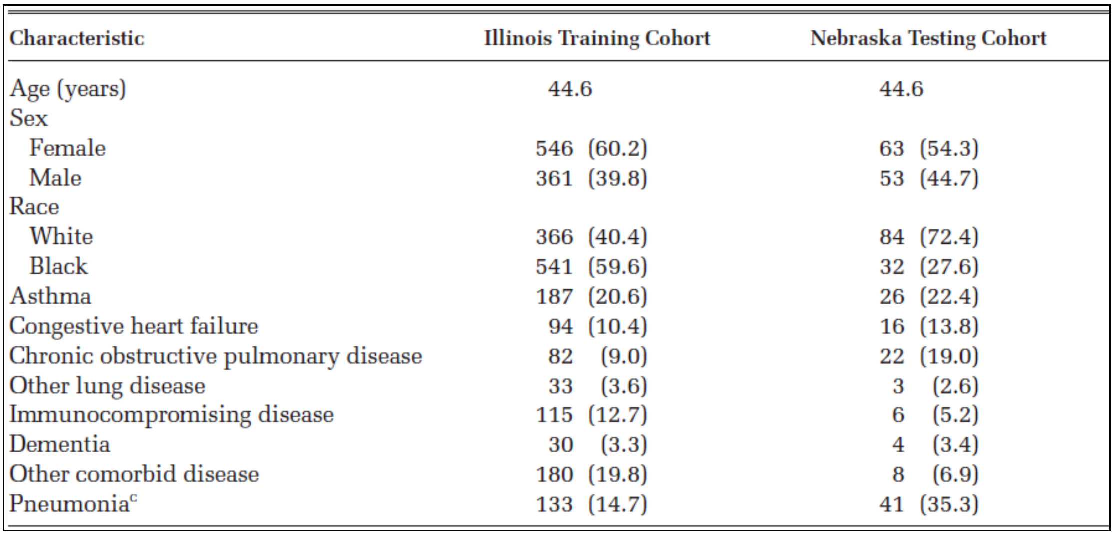


Figure 1: The accuracy from MLP and RBF



Characteristic	Illinois Training Cohort	Nebraska Testing Cohort
Age (years)	44.6	44.6
Sex		
Female	546 (60.2)	63 (54.3)
Male	361 (39.8)	53 (44.7)
Race		
White	366 (40.4)	84 (72.4)
Black	541 (59.6)	32 (27.6)
Asthma	187 (20.6)	26 (22.4)
Congestive heart failure	94 (10.4)	16 (13.8)
Chronic obstructive pulmonary disease	82 (9.0)	22 (19.0)
Other lung disease	33 (3.6)	3 (2.6)
Immunocompromising disease	115 (12.7)	6 (5.2)
Dementia	30 (3.3)	4 (3.4)
Other comorbid disease	180 (19.8)	8 (6.9)
Pneumonia[c]	133 (14.7)	41 (35.3)

Figure 2: Data set from the Illinois training set and the Nebraska test set

Figure 1 illustrates the number of hidden nodes from this paper. The solid line displays the number of hidden nodes and the accuracy percentage from using MLP algorithm and dash line shows the number of hidden nodes from using RBF algorithm. From the experiment, the architecture of both MLP and RBF networks was optimized by varying the number of hidden nodes from 2 to 74 nodes. The optimum number of hidden layer for MLP and RBF are 14 and 19, consequently. The accuracy is 89% for MLP and 86% for RBF. Paul S. Heckerling, Ben S. Gerber, Thomas G. Tape and Robert S. Wigton [13] present the results from predicting pneumonia disease using artificial neural network.

Figure 2 shows the training and testing data set. The training data set consists of 907 records from Illinois University and 116 records for the testing set from Nebraska University. The data set was collected from the Intensive Care Unit (ICU). Each record consists of 35 attributes and 2 classes, Pneumonia and non-Pneumonia. Figure 2 consists of three columns. The first column is an attribute name such as age, sex, race, asthma, congestive heart failure, chronic obstructive pulmonary disease, other lung disease, immunocompromising disease, dementia, other comorbid disease and pneumonia. The second column shows number and percentage of each attribute from the Illinois data set which we can read as follows:

The data set obtains 546 records for female patients which it is 60.2% of 907 records and 361 records for male patients which are equal to 39.8% of 907 records. There are 187 records for patients who have asthma symptom which it is 20.6% of 907 patient records.

The third column is similar to the second column but this column is the data set from the Nebraska Hospital. There are 63 records for female patients and 53 records for male patients. The percentage is 54.3 and 44.7, consequently.

They examined the network with 0, 1 and 2 hidden layer architectures. Finally, the architecture of network is 35-2-2. The learning rate and momentum are 0.5 and 0, respectively. The activation function employed is binary sigmoid transfer function. The artificial neural network can forecast correctly to 95%.

Ahmed M. Badawi, Manal Abdel Wahed and Shaimaa M. Elembaby [14] proposed the artificial neural network technique as the measurement to classify patients for Lower Urinary Tract Symptoms (LUTS) and Bladder Outlet Obstruction (BOO). There are 457 records. Each record consists of 4 different attributes. There are three different classes. The original data set was set into two groups: train set and test set. 300 records were used as the train set and 157 were used as the test set. The architecture of network is 4-11-3, 4 is the number of input nodes, 11 is the number of hidden nodes and 3 is the number of output nodes. The learning rate and momentum are set to 0.7 and 0.7. The accuracy percentage is 60.5%

Wen-wei Ouyang, Xiao-zhong Lin, Yi Ren, Yi Luo, Yun-tao Liu, Jia-min Yuan, Ai-hua Ou and Guo-zheng Li [15] proposed to classify the symptoms of TCM (Traditional Chinese Medicine) syndrome disease using naïve bayes algorithm. According to TCM, a body in harmony will not be diseased. The main syndromes are caused by dysfunction of the kidney, spleen, lungs and liver as reference in TCM literature. The symptoms that most often form the basis for TCM diagnoses can be TCM syndrome, For example, when the patients is feeling cold then it can actually be TCM syndrome like Spleen / Kidney Yang or Liver Qi stagnation or else like other syndromes.

The data set consists of 755 records, 384 records are used as the training set and 371 records is used as the testing set. Each record consists of 69 different symptoms and signs such as gender, age, height, weight, living environment, education, coughing, difficulty breathing, cardiac symptoms, chills and fever symptoms, upper respiratory tract symptoms, pale skin, head and body limbs, diet and stool. The rate of predicted accuracy is 78.55% and the error rate is 21.45%.

Susan P. Imberman Ph.D., Irene Ludwig, M.D., and Sarah Zelikovitz, Ph.D. [16] proposed to classify the group of esotropic patients by applying the decision tree technique. The patients in this study have the abnormal vision with various symptoms and treatment levels. The treatments may begin with taking medicines up to performing surgery depending on their signs and symptoms. The data set consists of 1307 patient records with an average of 14 visits per patient and also with the frequency of visits at most 52 visits and 2 visit at least. Each record consists of 54 different attributes. The main focus of this research is eyes deterioration. The accuracy using decision tree technique is 89% .

2.2 Background

In this study, the data set is collected from the Internal Patient Department (IPD) of Eyes, Ears, Nose, and Throat (EENT), faculty of medicine at Ramathibodi Hospital, Thailand. All patients came from all areas over the country.

The data set does not include the child patients who are lower than 15 years old because these patients cannot explain the exact symptoms to the clinician. The correct signs and symptoms are important for the correct diagnosis. Therefore, only adult patients are considered in this study.

The data set was collected from October 2008 to March 2010. There are 257 patient records. Each record consists of 35 attributes. The last attribute is a class attributes (disease) which consists of six different classes. These two following topics will be discussed in detail:

- Ear anatomy introductory
- Data set description

2.2.1 Ear anatomy introductory

The ear is a body part, which acts as a receiver for sound. The ear changes sound waves from outside into the signal of nerve impulses sent to the brain. Besides hearing, the ear plays a major role in the sense of balance and body position. Ear disease means any abnormal things that happen to the hearing system. The hearing system is very complicated, see figure 3 for the ear anatomy.

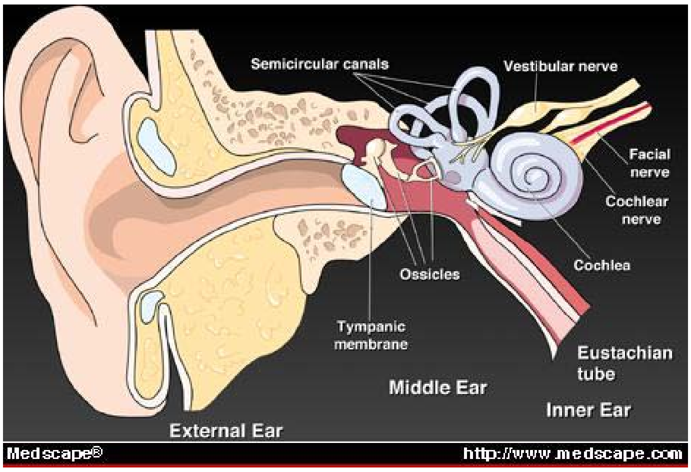


Figure 3: The ear anatomy [15]

The ear consists of 3 major parts: external ear, middle ear and inner ear [16]. The external ear, including the ear pinna, receives sound and passes through the middle ear. Such a sound is caused the eardrum membrane vibration to three bones (malleus or hammer, incus or anvil, and stapes or stirrup) in the middle ear. The sound wave transforms to nerve wave and passes through the inner ear cochlea, which then passes through the brain to translate into understanding and response to the sound.

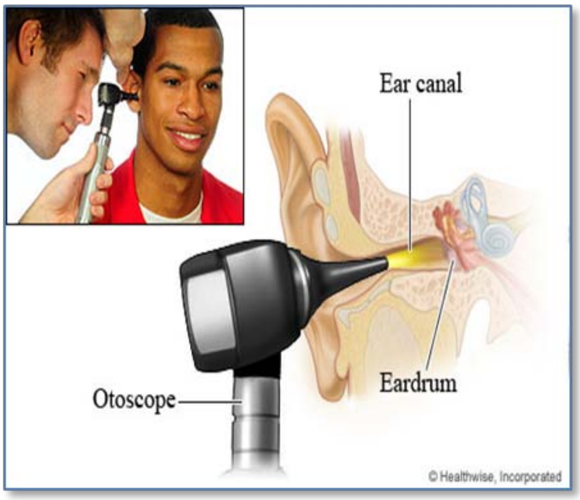


Figure 4: Position of Otoscope for ear examination [17]

In general, patients in IPD must go through the basic diagnosis examination from the Out Patient Department (OPD). Figure 4 shows the Otoscope which is an ear examination tool. The doctor always carefully checks for possible symptoms that may severely effect to hearing loss. When a patient is diagnosed with hearing, then the immediately surgery is needed. The clinicians must combine all their skills, knowledge and experiences to find out the cause of symptoms and advise the correct and effective treatments for patients.

2.2.2 Data set Description

In this study, there are 34 chief complaints (attributes) and 6 disease classes as following:

34 Chief Complaints consists of [18]

- Sex: male or female (Nominal, Yes or No)
- Age: (Ordinal)
- Otorrhea: purulent discharge (Nominal, Yes or No)
- Speaking: abnormal speaking of patient (Nominal, Loud or Low)
- Tinnitus: a sensation of noise as ringing or roaring (Nominal, Yes or No)
- AB Gap: air-bone gap is the difference between the threshold for hearing acuity by bone conduction and by air conduction (Nominal, Yes or No)
- Hearing Loss: a case where speaker can or cannot hear the sound (Nominal, Yes or No)
- Ototoxicity: having a harmful effect on the organs or nerves concerned with hearing and balance (Nominal, Yes or No)
- Noise Induce: live in loud pollution area (Nominal, Yes or No)
- Endocrine Induce: the patient has a disease that effects the ear problem such as hypothyroid or SLE (Nominal, Yes or No)
- Articulatory Defect: the performance of non-verbal (Nominal, Yes or No)
- Vertigo: to be in a whirl (Nominal, Yes or No)
- Fluid Fill Level High: the level of fluid in the middle ear is higher than normal (Nominal, Yes or No)
- Otogia: earache (Nominal, Yes or No)
- Facial Palsy: patient's face is not balance, effect from a dysfunction of the cranial nerve (Nominal, Yes or No)
- Keratin Mass: white keratin in the middle ear (Nominal, Yes or No)

- Ear Odor: terrible smell from the patient’s ear (Nominal, Yes or No)
- Fullness: the feeling of ears clogged (Nominal, Yes or No)
- Trauma: ear wound from the violent or the accident (Nominal, Yes or No)
- Finding Term: the period of ear symptom that has been beginning until present (Ordinal, 0-3 months or > 3 months)
- Heredity: the process by which particular traits or conditions are genetically transmitted from parents to offspring (Nominal, Yes or No)
- Deaf And Pregnancy: female patient who has ever been deaf some time during her pregnancy (Nominal, Yes or No)
- Infection: ear is infected (Nominal, Yes or No)
- Rhinitis: an inflammation of the nasal passages (Nominal, Yes or No)
- Fever: an abnormal elevation of the temperature of the body because of the disease (Nominal, Yes or No)
- Red Tympanic Membrane: ear drum is bulge and color changes to red (Nominal, Yes or No)
- Ruptured Eardrum: signs and symptoms of a ruptured eardrum may include (Nominal, Yes or No)
- Bulging: ear drum is bulgy (Nominal, Yes or No)
- Balance Problem: lost balance (Nominal, Yes or No)
- Dizziness: a sensation of faintness or an inability to maintain normal balance in a standing or seated position (Nominal, Yes or No)
- Headache: a pain in the head (Nominal, Yes or No)
- Back Ear Bulgy: a bulge at the back of external ear (Nominal, Yes or No)
- Retracted Eardrum: the Eustachian tube open abnormally which equalize the pressures in the middle ear from negative pressure in ear tube behind the ear drum. This causes the drum to become retracted (Nominal, Yes or No)
- Confusion: a mental state characterized by disorientation regarding to time, place or person causing bewilderment (Nominal, Yes or No)

6 Disease classes[19] consist of

- Acute Serous Otitis Media (ASOM)

ASOM is the kind of disease that happen when middle ear get inflammation with major symptoms include ear pain, fever, less hearing. Endoscope examination is required to diagnose the eardrum disease. Antibiotics and analgesics treatment is applied when symptoms are getting better, then surgery can be done afterward.

- Chronic Serous Otitis Media (CSOM)

CSOM is the disease of middle ear inflammation with major chronic ear pain, no fever, less hearing. Sometimes, there are complex acute symptoms. Endoscope examination is required to diagnose the disease as well as eardrum state. If there is an acute ear inflammation complex, the treatment can be done by antibiotics. After the recovery, ear can be treated by surgery.

- Otosclerosis

A genetic disease of bone changes and destroys the middle ear bone. The important symptoms are slowly decreased hearing. To reduce patients bone destruction and abnormality, firstly the treatment is done by giving drugs. Finally, the hearing aid or surgery is considered, depending on patients' choice.

- Cholesteatoma

The state has accumulated Keratinizing Epithelium in each middle ear position. The formation of Keratin causes the destroying bone of middle ear by enzyme which is led to hearing loss symptom. It needs to be treated by incision in order to clot out Cholesteatoma.

- Conductive Hearing Loss

The hearing loss of impaired voice disorders due to several reasons such as auditory canal constricts and cerumen (the yellow or brown waxy secretions that are produced by sweat glands in the external ear canal) and inflammation. Common symptoms consist of decreased hearing, liquid and purulent flowing out of the ear. Surgery is the way to cure such a disease.

- Sensorineural Hearing Loss

This happens from many causes such as infection, head impact from accidents, drugs and chemicals. Common symptoms consist of decreased hearing, earache, fever, headache, dizziness. Treatment for the disease is done by sterilization drug or cochlear implant.

CLASS	FREQUENCY	PERCENTAGE (%)
Acute Serous Otitis Media	37	14.40
Chronic Serous Otitis Media	108	42.02
Otosclerosis	41	15.95
Cholesteatoma	28	10.89
Sensorineural Hearing Loss	19	7.39
Conductive Hearing Loss	24	9.35
	257	100

Table 1: Frequency and percentage ratio of each class

Six different classes of disease are shown in Table 1. The first column displays the disease classes. The second column shows frequency or number of records in each class. The last column is the percentage ratio of each class. There are 37 records for Acute Serous Otitis Media, 108 records for Chronic Serous Otitis Media, 41 records for Otosclerosis, 28 records for Cholesteatoma, 19 records for Sensorineural Hearing Loss and 24 records for Conductive Hearing Loss.

Next chapter discusses the methodology employed in the experiments.

CHAPTER 3
METHODOLOGY

In this study, the intelligent capability of data mining is used to extract hidden patterns in the data set, as can be seen in the following figure.

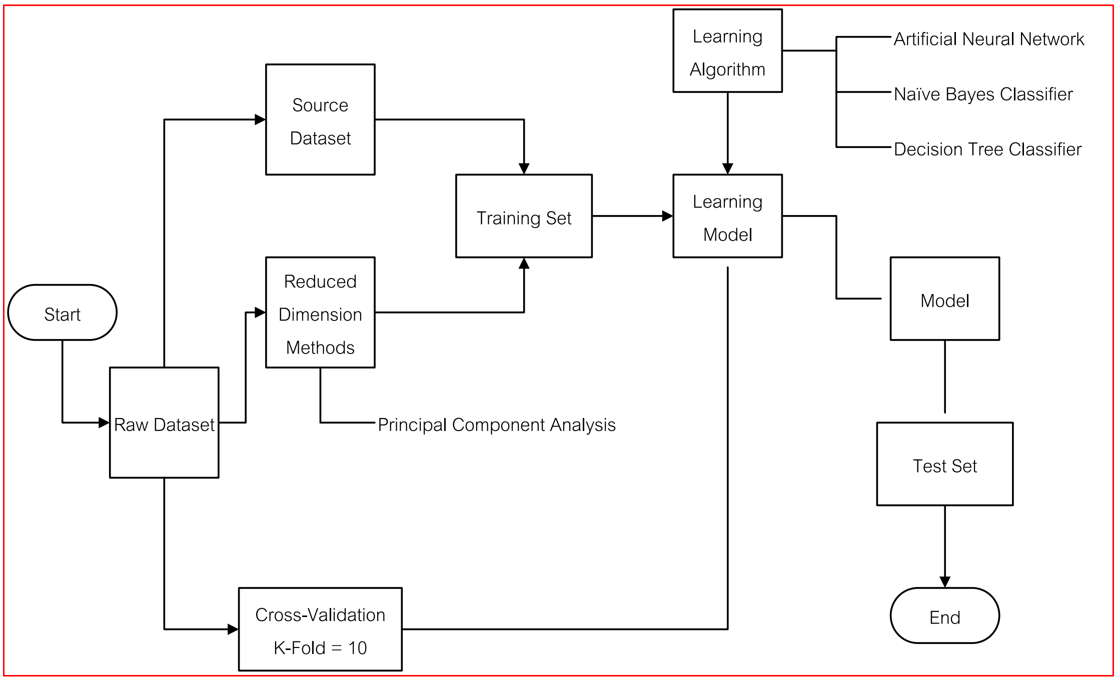


Figure 5: Data processing

Figure 5 shows the experimental processes. It starts from raw data set which consists of 35 attributes, 257 records and 6 different classes. The raw data set is transformed using Min-max normalization. The details of transformation will be discussed later. Here, the raw data set after the transformation process will be called as a data set. The data set can be employed in three different ways. Firstly, the data set is used as the training set. This training set will be experimented using three different learning algorithms 1) Artificial Neural Network 2) Naïve Bayes and 3) Decision Tree. Then, the performances of these algorithms are measured using the first data set. Secondly, the data set is processed by the reduced dimension method; PCA for this study. Numbers of input attributes are reduced. These reduced attributes will be used as the input to train these three learning algorithms. After the training process with reduced attributes, the performance of three algorithms is measured. Lastly, the data set using cross-validation will be trained with these three algorithms and then the performance of models is measured. Here, the cross-validation is set to 10 folds.

Generally, the training set is employed for building the classification model. The test set is utilized for measuring the accuracy of the model

The raw data set is transformed using normalization process which is the process to convert the raw data set into the range between 0.0 and 1.0

There are many techniques for data transformation. Min-Max normalization is one of widely used techniques. It is based on linear function. The result is formed as a fraction, which is less than 1.0. The transformed values can be computed as the following equation:

$$v'_i = \frac{v_i - min_i}{max_i - min_i}, \qquad \ldots [20]$$

Where:

V'_I = new value of V_i

min_i = minimum of V attribute

max_i = maximum of V attribute

V_i = the value of V attribute

Three classifiers artificial neural network, Naïve Bayes, and Decision tree are used as learning algorithms. The discussion of each technique is as following:

3.1 Artificial Neural Network (ANN)

The ANN is a famous technique, which simulates biological human brain working process. Human brain consists of neurons, synapses, dendrite, and axons. The ANN consists of input layer, hidden layer, and output layer. Each layer consists of many neurons which are completely connected as network.

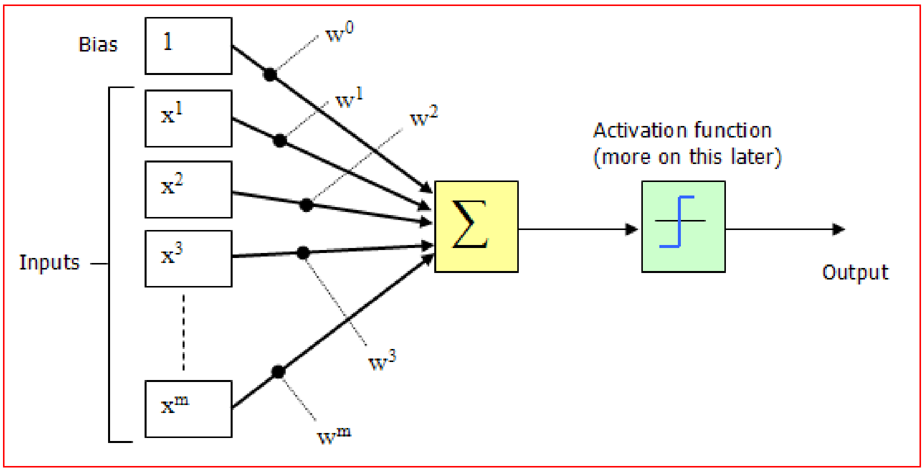


Figure 6: Perceptron [21]

Where:

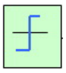
$X^1 \ldots X^m$	are	input nodes
$W^1 \ldots W^m$	are	connection weights
Bias	is	another input node, the value of bias is 1
$\sum$	is	the summation processor
	is	the activation function

Figure 6 illustrates a perceptron. A perceptron or a single layer perceptron [22] is a supervised classification algorithm. This single layer perceptron is a type of linear classifier which is a simple kind of feedforward neural network.

The perceptron was invented in 1957 by Frank Ronsenblatt [23]. The perceptron learning rule is a method for finding weights in a network. In general, the assumptions of the learning rule include:

1) a one layer network is used with a binary step function
2) mean square error (MSE) is used as a cost function
3) target values are between + 1 and – 1

However, there are some limitations with this network, for example, perceptron cannot solve the XOR (exclusive OR) problem because a single layer generates only a linear decision boundary.

In 1969, Minsky and Papert [24] proposed the solution to XOR problem by using a second layer of units. This layer combines the response from perceptron units.

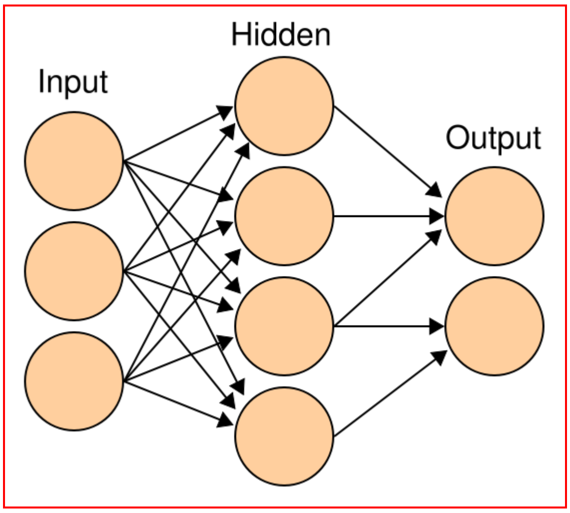


Figure 7: Multi-layer perceptron [25]

Figure 7 shows the multi-layer perceptron or multi-layer network. Such a figure illustrates 3 nodes in the input layer, 4 nodes in the hidden layer, and 2 nodes in the output layer.

Input layer: each node represents each input attribute from the data set.

Output layer: each node denotes each output class.

Hidden layer: one or more layer between the input and output layer. It is claimed that the hidden layer acts as the function approximator that can approximate any function which is so-called “Universal Approximator” [26].

Here are some important properties of this architecture.

1) layers are fully connected
2) there are no direct connections between input and output layer
3) there are no connections within the same layer
4) input, hidden and output units do not have to be the same numbers.
5) bias is often included as an extra weights

The purpose of the different layers are: The first layer draws linear boundary. The second layer combines the boundary. The third layer generates arbitrary boundary shapes [27]

Backpropagation (BP) is a learning algorithm for the feed forward neural network. BP was inventes by Rumelhart, Hinton and Williams (1986) [28]. This learning algorithm has 2 phrases. In the first phrase (forward pass), input signals are feed-forwarded or propagated through the network. For the second phrases (backward pass). The error is propagated starting from the output layer. The error is the difference between actual (computed) output values and target (desired) output values.

In this study, BP has been utilized for training the network. BP learning algorithm can be summarized as follows:

```
1) the learning rate ($\eta$), momentum ($\alpha$) are set
2) initial random weights and bias are set
3) repeat until criteria is satisfied (number of iterations (epoch) or error are met)
       input patterns are presented to input layer
       function signal for input, hidden and output units are computed
       target is presented to output units
       error signals for output units are computed
       error signals for hidden units are computed
       all weights are updated
       input next patterns and targets are presented
   end repeat [29].
```

Each full presentation of all patterns is called epoch. The order of presented training patterns prefers to be random to avoid the correlation between consecutive training pairs. Choosing initial weights are crucial because, these initial weights will determine the starting points from the global minimun. The learning rate (η) controls the step size when weights are adjusted. With small learning rate, the learning is smooth but slow. Whereas, the big learning rate can speed the learning but there is a risk of divergence. The purpose of momentum (α) is to reduce instablility problem while increasing th convergent rate. Rate of change (E) can be used as the stopping criteria usually it is set as a small value. Nevertheless, the goal is to classify the new pattern correctly.

3.2 Naïve Bayes (NB)

NB is a probalistic based on Naïve Bayes' Theorem. This classifier assumes feature independency. This means that the existence or non-existence of a particular feature of a class is not related to any other features.

The class-conditional independence assumption greatly simplifies the training step since the one-dimensional class-conditional density for each feature can be estimated individually. While, the class-conditional independence between features is not true in general. Research [30] shows that this optimistic assumption works well in practice. This assumption of class independence allows the Naive Bayes classifier to better estimate the parameters required for accurate classification while using less training data than many other classifiers [31]. This makes it particularly effective for data sets containing many predictors or features [32].

Bayes's classifier is trained under supervised learning. The conditional probability under Bayes' theorem can be shown as following equation: [33]

$$p(C|F_1, \dots , F_n) = \frac{p(C)p(F_1,\dots,F_n|C)}{p(F_1,\dots,F_n)}$$

When C is a class variable

$F_1,\dots,F_n$ are feature variables

It is important to note that, NB classifier combines the above model with a common rule. The common rule is to pick the hypothesis that is the most probable.

3.3 Decision Tree (DT)

Decision tree is a non parametric statistics because assumptions about the distribution of the class variables are not required.

Several popular decision tree algorithms include ID3 [34], C4.5 and CART (Classification and regression tree). The decision tree algorithm is iterative and based on Hunt algorithm. Local minimum without backtracking is employed. Normally, the result may not be optimum but it is considered to be very fast.

Decision tree is a tree-like graph. The general idea of building the tree is starting with root node and recursively add child nodes until the training data is fit. Each branch in the tree represents the decision rule. Each leaf node is the predicted value for that node. In other words, the leaf node is used to classify the instances.

The interest question is how to select the attribute. A feature is selected as the split attribute according to the impurity. Most well known indices to measure degree of impurity are entropy, GINI index and classification error. These indices measure how well an attribute separate the training samples to the target classes. Here, GINI index is applied in the experiments.

GINI index is calculated by summing the product of the probability of each chosen item with the probability of a mistake in classifying that item. If all cases fall in the same category, the GINI index will be minimum value (zero) [35].

Here is the formula to compute GINI index [36].

.

$$I_G(f) = \sum_{i=1}^{m} f_{,}(1 - f_i)$$

There are several advantages in employing the decision tree such as 1) easy to understand and interpret 2) able to handle both numerical and categorical data 3) require little data preparation.

3.4 Principal Component Analysis (PCA)

Since there are a large number of input attributes, it is more likely that subsets of these variables might be correlated. These correlated variables might reduce the accuracy and reliability of a classification model because of the overfitting problem. In addition, these redundant variables can increase the processing costs. Thus, reducing the dimensionality without losing the information or performance accuracy is one of the key ideas in mining the data.

PCA was invented by Karl Pearson [37] in 1901. The procedure includes orthogonal linear transformation, in which, a set of possible corrected variables are converted to a new co-ordinate system. The greatest variance by the projection lies on the first co-ordinate.

General algorithm for PCA is as follows:

1) obtain the data
2) subtract the mean
3) calculate variance
4) calculate eigenvectors and eigenvalues of co-variance
5) choose components and finding a feature vector
6) develop the new data set

3.5 Cross-Validation

Cross-Validation is a data mining technique which aims to increases the accuracy percentage of the result by dividing the original data set into k-groups or k-folds. In general, the number of partition is set to 10.

The objective of k-fold cross-validation is increasing the accuracy of the prediction. The algorithm selects one partition as the test set and other partitions as the training set. All partitions are used as the test set. Then, the learning step will be stop. The error from each round will be recorded. Finally, the numbers of results is equal to numbers of partitions. The results are then summed and divided by number of folds (k).

The experimental results are illustrated in next chapter.

CHAPTER 4
EXPERIMENTAL RESULT

This chapter illustrates the experiments and their results. All experiments are simulated using Matlab [38], the suitable program for calculation in the form of matrix. The experiments can be divided into 3 main categories as following:

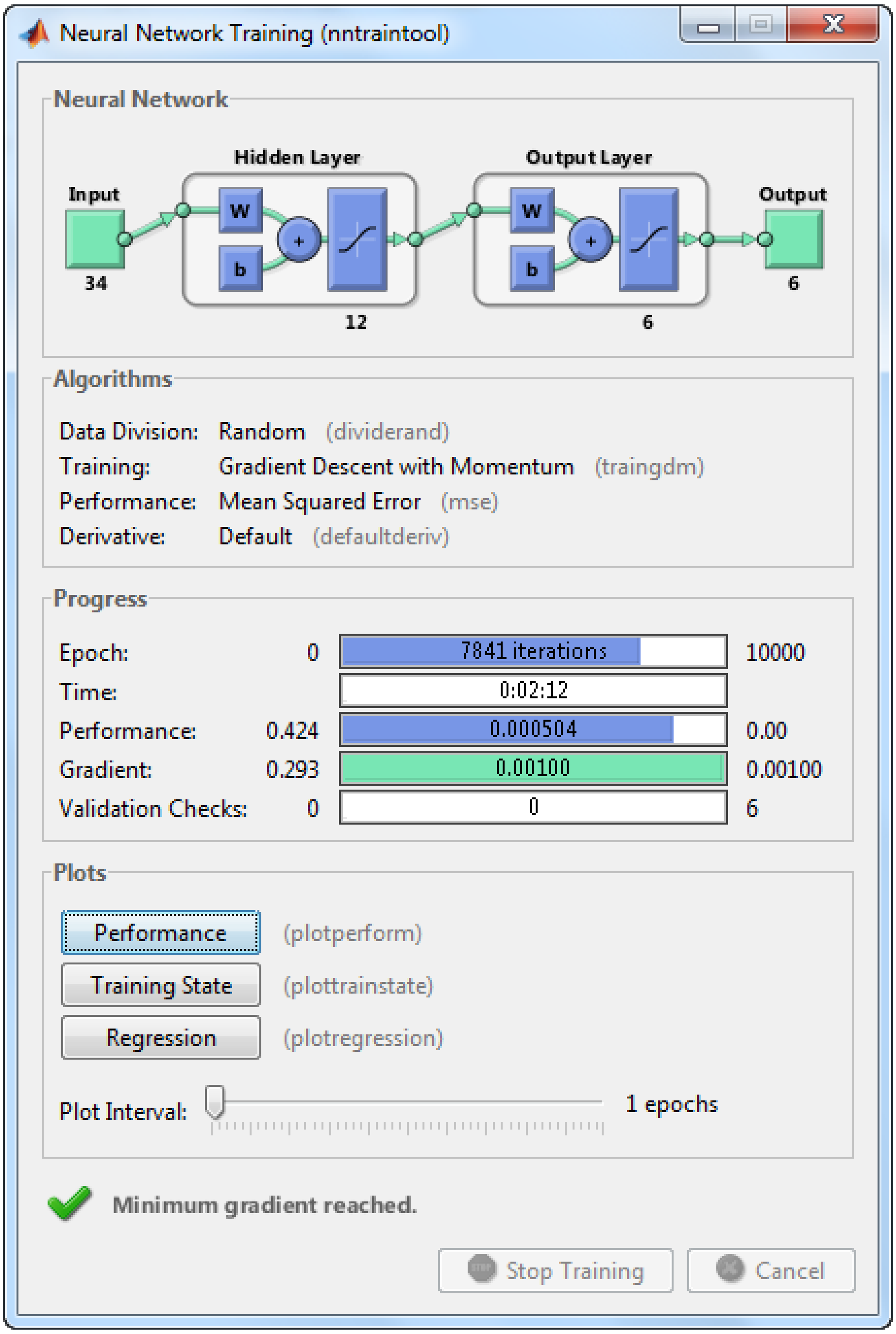
- Experiments without the Principal component analysis
    - Training with Artificial neural network
    - Training with Naïve Bayes
    - Training with Decision tree
- Experiments with the Principal component analysis (PCA)
    - Training with Artificial neural network
    - Training with Naïve bayes
    - Training with Decision tree
- Experiments with K-fold cross-validation
    - Training with Artificial neural network
    - Training with Naïve Bayes
    - Training with Decision tree

4.1 Experiments without the Principal component analysis (PCA)

The ratio of the train set and test set is 60:40 for all techniques. Therefore, 257 patient records are divided into 154 records for training and 103 records for testing.

4.1.1 Training with Artificial neural network

The multilayer perceptron is trained under supervision using the back-propagation algorithm [39] [40]. Both learning rate and momentum value are set to 0.1. Note that, 0.1 is the number that came from the trial experiments. With this number the optimal results is produced.



Algorithms

Data Division: Random (dividerand)
Training: Gradient Descent with Momentum (traingdm)
Performance: Mean Squared Error (mse)
Derivative: Default (defaultderiv)

Progress

Epoch:	0	7841 iterations	10000
Time:		0:02:12	
Performance:	0.424	0.000504	0.00
Gradient:	0.293	0.00100	0.00100
Validation Checks:	0	0	6

(plotperform)
(plottrainstate)
(plotregression)

Plot Interval: 1 epochs

✓ Minimum gradient reached.

Figure 8: Neural network is trained on data set without PCA

Figure 8 illustrates the result from training set using Matlab. This result consists of 4 parts; neural network, algorithms, progress and plots.

First part, the neural network, displays the architecture of the network. From the figure, there are 34 nodes, 12 hidden nodes with sigmoid activation function and 6 output nodes with sigmoid activation function.

In the second part, algorithms, shows the details of the training algorithm. These include data division is random. The training utilizes gradient descent with momentum. The performance is measured by mean square error.

Third part, progress, displays number of epoch, time performance (MSE) and gradient error.

Last part, plots, shows the result in the form of 2-D graph.

		Prediction result					
		A	B	C	D	E	F
Diagnosis result	A	15	0	0	0	0	0
	B	0	11	0	0	0	0
	C	0	0	43	0	0	0
	D	0	0	0	10	0	0
	E	0	0	0	0	17	0
	F	0	0	0	0	0	7

Table 2: Confusion matrix of neural network is trained on data set without PCA

Table 2 displays the confusion matrix from the experiment using artificial neural network. The confusion matrix explains the information on the actual diagnostic and the predicted class from the network. In the table, each row represents the actual diagnostics, whereas, each column of the matrix represents the predicted results from the network. The interpretation is as follows:

The labels A, B, C, D, E and F represent ear disease classes. For example, A is ASOM, B is Cholesteatoma, C is CSOM, D is Conductive hearing loss, E is Otosclerosis and F is Sensorineural hearing loss. The performance of the neural network approach can be evaluated using the number in the matrix.

- Horizontal reading from diagnostic result for ear disease class A (Acute Serous Otitis Media) is 15 and the predicted result from the network is also 15 which means all records are predicted correctly. Thus, the accuracy percentage for class A is 100%

- Horizontal reading from diagnostic results for ear disease class B (Cholesteatoma) is 11 and the predicted result from the network is also 11. All records are predicted correctly. Thus, the accuracy percentage for class B is 100%.

- Horizontal reading from diagnostic results for ear disease class C (Chronic Serous Otitis Media) is 43 and the predicted result from the network is also 43. All records are predicted correctly. Thus, the accuracy percentage for class C is 100%.

- Horizontal reading from diagnostic results for ear disease class D (Conductive Hearing Loss) is 10 and the predicted result from the network is also 10. All records are predicted correctly. Thus, the accuracy percentage for class D is 100%.

- Horizontal reading from diagnostic results for ear disease class E (Otosclerosis) is 17 and the predicted result from the network is also 17. All records are predicted correctly. Thus, the accuracy percentage is for class E 100%.

- Horizontal reading from diagnostic results for ear disease class F (Sensorineural Hearing Loss) is 7 and the predicted result from the network is also 7. All records are predicted correctly. Thus, the accuracy percentage for class F is 100%.

Finally, the architecture of the network in this study is 35-12-6.

4.1.2 Training with Naïve Bayes

		Prediction result					
		A	B	C	D	E	F
Diagnosis result	A	15	0	0	0	0	0
	B	0	11	0	0	0	0
	C	0	0	43	0	0	0
	D	0	0	0	10	0	0
	E	0	0	0	0	17	0
	F	0	0	0	0	0	7

Table 3: Confusion matrix of Naïve Bayes is trained on data set without PCA.

The predicted output using naïve bayes classifier can be illustrated as confusion matrix in Table 3. The confusion matrix demonstrates the number of actual diagnostic class and the predicted class. Each row of the matrix represents the actual diagnostics, whereas, each column of the matrix represents the predicted results. The results can be interpreted as follows

The labels A, B, C, D, E and F denote ear disease classes. For example, A represents Acute Serous Otitis Media, B is Cholesteatoma, C is Chronic Serous Otitis Media, D is Conductive Hearing Loss, E is Otosclerosis and F is Sensorineural Hearing Loss. Performance of the Naïve bayes approach can be evaluated using numbers in the matrix. The examples of interpretations are:

- Horizontal reading from diagnostic results for ear disease class A (Acute Serous Otitis Media) is 15 and the predicted result from the network is also 15. All records are predicted correctly. Thus, the accuracy percentage of class A is 100%

- Horizontal reading from diagnostic results for ear disease class B (Cholesteatoma) is 11 and the predicted result from the network is also 11. All records are predicted correctly. Thus, the accuracy percentage of class B is 100%.

- Horizontal reading from diagnostic results for ear disease class C (Chronic Serous Otitis Media) is 43 and the predicted result from the network is also 43. All records are predicted correctly. Thus, the accuracy percentage of class C is 100%.

- Horizontal reading from diagnostic results for ear disease class D (Conductive Hearing Loss) is 10 and the predicted result from the network is also 10. All records are predicted correctly. Thus, the accuracy percentage of class D is 100%.

- Horizontal reading from diagnostic results for ear disease class E (Otosclerosis) is 17 and the predicted result from the network is also 17. All records are predicted correctly. Thus, the accuracy percentage of class E is 100%.

- Horizontal reading from diagnostic results for ear disease class F (Sensorineural Hearing Loss) is 7 and the predicted result from the network is also 7. All records are predicted correctly. Thus, the accuracy percentage of class F is 100%.

It can be seen that the network correctly classified 102 records from a total test data set 102 records. The accuracy percentage is 100 %.

4.1.3 Training with Decision tree

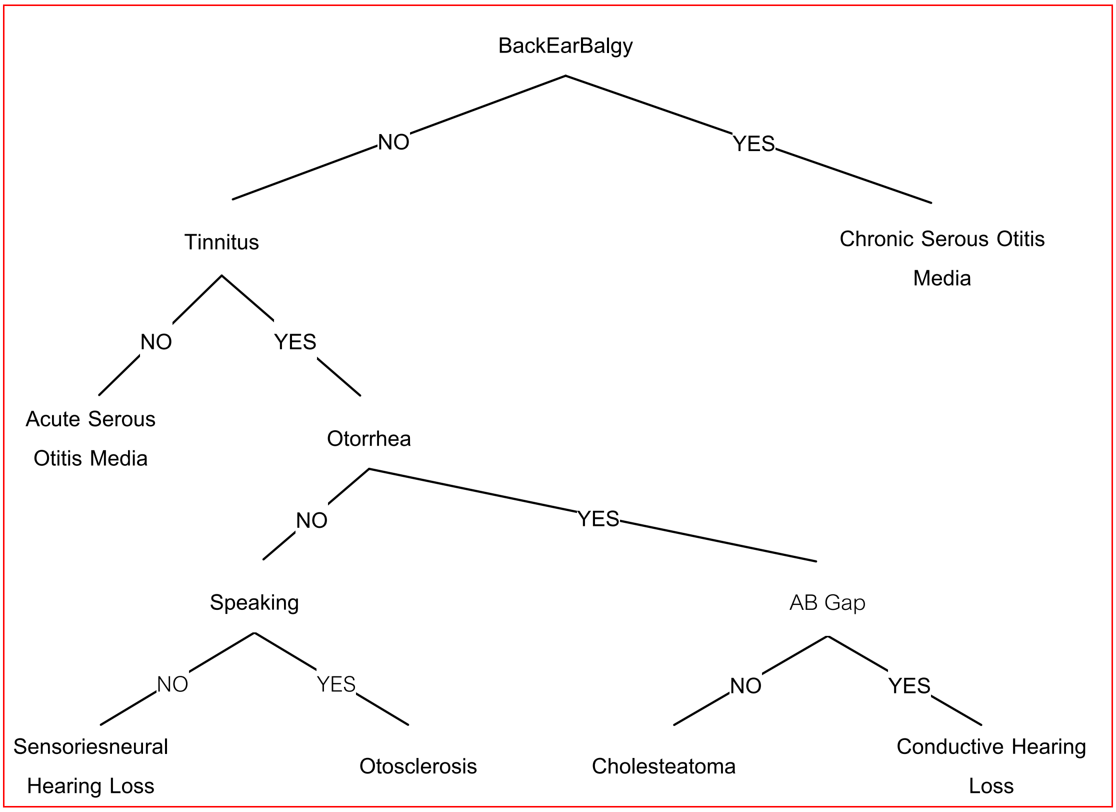


Figure 9: The result from training using decision tree on data set without PCA.

Figure 9 shows the result from the training using decision tree. We can go through the decision tree as top-down approach until the leaf node is reached. For example, on top of the graph, if the value of BackEarBalgy in the test set is Yes, the result from predicting is Chronic Serous Otitis Media but if the value of BackEarBalgy is No,the attribute on the left (Tinnnitus) is considered instead. When the Tinnitus value in the test set is No, the predicted result is Acute Serous Otitis Media but if the value of Tinnitus is Yes, Otorrhea is considered as the next attribute. Then when the Otorrhea value is Yes, AB Gap attribute is considered, if value of AB Gap attribute is Yes, the result from predicting is Conductive Hearing Loss, in the other hand, the result is Cholesteatoma. If the value of Otorrhea is No, Speaking attribute is considered. When the value of Speaking attribute is Yes, the result from predicting is Otosclerosis but if the value of Speaking attrubute is No, the predicted result is Sensoriesneural Hearing Loss.

The training using GINI split method can be discuss as follows:

Step one:

Attributes	GINI Split
BackEarBulgy	0.455197133
RetractedEardrum	0.455197133
Confussion	0.455197133
Tinnitus	0.533892852
Fever	0.533892852
Headache	0.538674991
Vertigo	0.557665689
Dizziness	0.562135262
Trauma	0.567131319
BalanceProblem	0.574809644
Infection	0.581274659
Fullness	0.581486552
FluidFillLevelHigh	0.585434995
Rhinitis	0.585434995
RedTempanicMembrane	0.585434995
Bulging	0.585434995
Otorrhea	0.587476113
FacialPulsy	0.590890937
FindingTerm	0.622120645
KeratinMass	0.631329775
EarOdor	0.631329775
AB_Gap	0.645252803
Ototoxicity	0.650015361
Speaking	0.659508234
Hearing_Loss	0.659508234
Articulatory_Defect	0.659508234
Otogia	0.659508234
DeafandPregnancy	0.67358448
NoiseInduce	0.691770902
EndocrineInduce	0.691770902
Heredity	0.69956408
Sex	0.726733324
Age	0.731252168
RuptureEardrum	0.751300728

Table 4: Result of GINI split in step one

Table 4 shows the GINI value calculated for each attribute. The values are shown in ascending order (lowest to highest). Generally, the best splitted attribute should be the attribute with the lowest GINI value. Therefore, attributes: BackEarBulgy, RetractedEardrum and Confussion can be choosen as the splitted attributes. Here, BackEarBulgy is selected as the best split for the first step.

Step two:

Attributes	GINI Split
Tinnitus	0.541956882
FluidFillLevelHigh	0.541956882
Rhinitis	0.541956882
Fever	0.541956882
RedTempanicMembrane	0.541956882
Bulging	0.541956882
Infection	0.563408692
BalanceProblem	0.564194458
Fullness	0.566113625
Trauma	0.566113625
Dizziness	0.567397586
Vertigo	0.568773449
FacialPulsy	0.568773449
Otorrhea	0.585218703
KeratinMass	0.597897898
EarOdor	0.597897898
Headache	0.599334357
AB_Gap	0.617251462
Ototoxicity	0.627259259
Speaking	0.637891738
Hearing_Loss	0.637891738
Articulatory_Defect	0.637891738
Otogia	0.637891738
FindingTerm	0.659986505
DeafandPregnancy	0.682336182
NoiseInduce	0.691327913
EndocrineInduce	0.691327913
Heredity	0.703882195
Sex	0.714583333
Age	0.735000000
RuptureEardrum	0.783950617
RetractedEardrum	0.783950617
Confussion	0.783950617

Table 5: Result of GINI split in step two

The coresponding GINI value for each attributed is recomputed. The results of GINI values for all attributes are shown in ascending order as Table 5. In generally, the splitted attribute is the one with the lowest GINI value. There are 6 attributes with the lowest value; Tinnitus, Fluidfilllevelhigh, Rhinitis, Fever, Redtempanicmembrane and Bulging. Here, Tinnitus attribute is chosen as the split attribute for the second step.

Step three:

Attributes	GINI Split
Otorrhea	0.464918650
Infection	0.464918650
Headache	0.472582144
KeratinMass	0.478782558
EarOdor	0.478782558
Fullness	0.478782558
Trauma	0.478782558
BalanceProblem	0.484108538
AB_Gap	0.502393692
Dizziness	0.502393692
Ototoxicity	0.516379640
Speaking	0.528629579
Hearing_Loss	0.528629579
Articulatory_Defect	0.528629579
Vertigo	0.528629579
Otogia	0.528629579
FacialPulsy	0.528629579
Sex	0.603866382
NoiseInduce	0.604098153
EndocrineInduce	0.604098153
DeafandPregnancy	0.613297151
Heredity	0.621393035
Age	0.660999814
FindingTerm	0.704668963
FluidFillLevelHigh	0.728001782
Rhinitis	0.728001782
Fever	0.728001782
RedTempanicMembrane	0.728001782
RuptureEardrum	0.728001782
Bulging	0.728001782
RetractedEardrum	0.728001782
Confussion	0.728001782

Table 6 Result of GINI split in step three

Table 6 shows the GINI values that are re-calculated for all the corresponding attributes. From the results in table 6, Otorrhea gives the lowest GINI value. Therefore, this attribute is chosen as the split attribute in the third step.

Step four:

Attributes	GINI Split
Speaking	0
Hearing_Loss	0
Articulatory_Defect	0
Vertigo	0
Otogia	0
FacialPulsy	0
BalanceProblem	0
Headache	0
NoiseInduce	0.186393290
EndocrineInduce	0.186393290
Sex	0.225225225
Heredity	0.225225225
DeafandPregnancy	0.337297297
Age	0.371968122
Ototoxicity	0.412912913
FindingTerm	0.434034034
Otorrhea	0.438276114
AB_Gap	0.438276114
FluidFillLevelHigh	0.438276114
KeratinMass	0.438276114
EarOdor	0.438276114
Fullness	0.438276114
Trauma	0.438276114
Infection	0.438276114
Rhinitis	0.438276114
Fever	0.438276114
RedTempanicMembrane	0.438276114
RuptureEardrum	0.438276114
Bulging	0.438276114
Dizziness	0.438276114
RetractedEardrum	0.438276114
Confussion	0.438276114

Table 7: Result of GINI split in step four

Table 7 shows the GINI values that are re-computed for their corresponding attributes. It can be seen that 8 attributes: Speaking, Hearing_loss, Ariculatory_Defect, Vertigo, FacialPulsy, BalanceProblem and Headache gave the lowest GINI value. Here, speaking is chosen as the split attribute in the forth step.

Step five:

Attributes	GINI Split
AB_Gap	0
Ototoxicity	0
KeratinMass	0
EarOdor	0
Fullness	0
Trauma	0
BalanceProblem	0
Dizziness	0
Headache	0
Age	0.311111111
FindingTerm	0.384000000
Sex	0.444444444
Otorrhea	0.497777778
Hearing_Loss	0.497777778
NoiseInduce	0.497777778
EndocrineInduce	0.497777778
Articulatory_Defect	0.497777778
Vertigo	0.497777778
FluidFillLevelHigh	0.497777778
Otogia	0.497777778
FacialPulsy	0.497777778
Heredity	0.497777778
DeafandPregnancy	0.497777778
Infection	0.497777778
Rhinitis	0.497777778
Fever	0.497777778
RedTempanicMembrane	0.497777778
RuptureEardrum	0.497777778
Bulging	0.497777778
RetractedEardrum	0.497777778
Confussion	0.497777778

Table 8: Result of GINI split in step five

Table 8 shows the GINI value in ascending order for each attribute. Here AB_Gap attribute is chosen as the split attribute in the fifth step.

		Prediction result					
		A	B	C	D	E	F
Diagnosis result	A	15	0	0	0	0	0
	B	0	11	0	0	0	0
	C	0	0	43	0	0	0
	D	0	0	0	10	0	0
	E	0	0	0	0	17	0
	F	0	0	0	0	0	7

Table 9: Confusion matrix of decision tree is trained on data set without PCA.

Table 9 illustrates the confusion matrix from decision tree. The result can be interpreted as follows:

The labels A, B, C, D, E and F denote ear disease classes. For example, A represents Acute Serous Otitis Media, B is Cholesteatoma, C is Chronic Serous Otitis Media, D is Conductive Hearing Loss, E is Otosclerosis and F is Sensorineural Hearing Loss. Performance of the Naïve Bayes approach can be evaluated using numbers in the matrix. The examples of interpretations are:

- Horizontal reading from diagnostic results for ear disease class A (Acute Serous Otitis Media) is 15 and the predicted result from the network is also 15. All records are predicted correctly. Thus, the accuracy percentage of class A is 100%

- Horizontal reading from diagnostic results for ear disease class B (Cholesteatoma) is 11 and the predicted result from the network is also 11. All records are predicted correctly. Thus, the accuracy percentage of class B is 100%.

- Horizontal reading from diagnostic results for ear disease class C (Chronic Serous Otitis Media) is 43 and the predicted result from the network is also 43. All records are predicted correctly. Thus, the accuracy percentage of class C is 100%.

- Horizontal reading from diagnostic results for ear disease class D (Conductive Hearing Loss) is 10 and the predicted result from the network is also 10. All records are predicted correctly. Thus, the accuracy percentage of class D is 100%.

- Horizontal reading from diagnostic results for ear disease class E (Otosclerosis) is 17 and the predicted result from the network is also 17. All records are predicted correctly. Thus, the accuracy percentage of class E is 100%.

- Horizontal reading from diagnostic results for ear disease class F (Sensorineural Hearing Loss) is 7 and the predicted result from the network is also 7. All records are predicted correctly. Thus, the accuracy percentage of class F is 100%.

It can be seen that the network correctly classified 102 records from a total test data set 102 records. The accuracy percentage is 100 %.

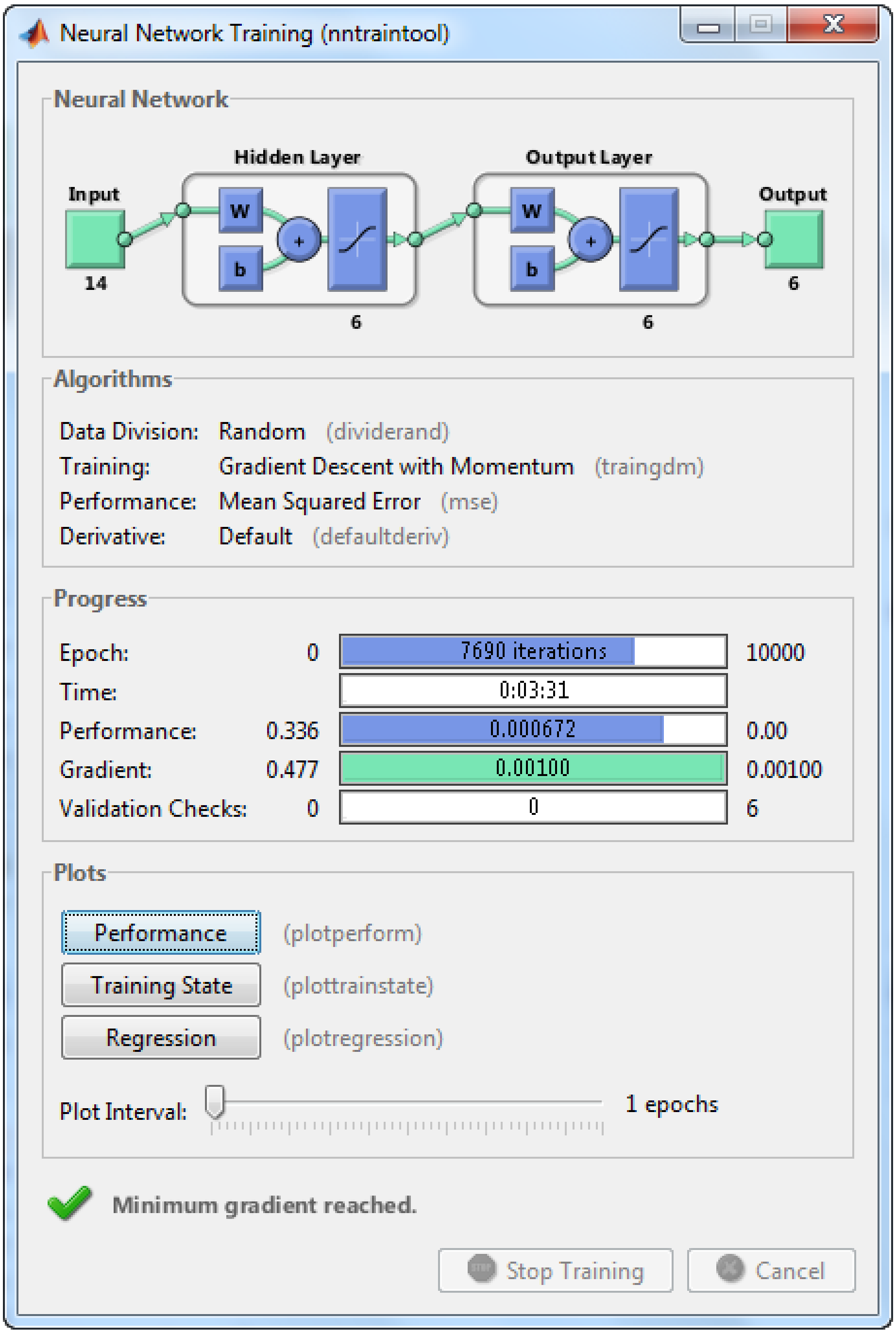
4.2 Experiments with Principal component analysis (PCA)

The PCA technique has been utilized to reduce number of input attributes. The reduction details can be seen in Appendix C. After the PCA process, there are 14 attributes, namely AB Gap, BackEarBulgy, BalanceProblem, Confussion, FacialPulsy, Fever, Fullness, Hearing Loss, Otogia, Otorrhea, Ototoxicity, RetractedEardrum, Speaking and Vertigo are employed as input attributes. In other words, these attributes are main symptoms used as input.

The data set used in all techniques (Artificial neural network, Naïve bayes, and Decision tree) is divided into 2 groups: train set and test set. The ratio is 60:40.

4.2.1 Training with Artificial neural network with PCA

The ANN is simulated using Matlab. The learning rate and momentum of training the network is set to 0.1. The value 0.1 is obtained from experimental results that are expected to give the optimal result.



Algorithms

Data Division: Random (dividerand)
Training: Gradient Descent with Momentum (traingdm)
Performance: Mean Squared Error (mse)
Derivative: Default (defaultderiv)

Progress

Epoch:	0	7690 iterations	10000
Time:		0:03:31	
Performance:	0.336	0.000672	0.00
Gradient:	0.477	0.00100	0.00100
Validation Checks:	0	0	6

(plotperform)
(plottrainstate)
(plotregression)

Plot Interval: 1 epochs

✓ Minimum gradient reached.

Figure 10: Neural network is trained on data set with PCA.

Figure 10 shows the result of training the network using Matlab. Here, the network architecture is 15-6-6. The first number 15 is the number of input nodes which is 14 including class attribute. The second number 6 is the number of hidden nodes. The third number 6 is the number of output nodes or number of classes.

		Prediction result					
		A	B	C	D	E	F
Diagnosis result	A	15	0	0	0	0	0
	B	0	11	0	0	0	0
	C	0	0	43	0	0	0
	D	0	0	0	10	0	0
	E	0	0	0	0	17	0
	F	0	0	0	0	0	7

Table 10: Confusion matrix of neural network is trained on data set with PCA.

The confusion matrix in table 10 prediction result by the artificial neural network. The discussion is as follows:

- Horizontal reading from diagnostic results for ear disease class A (Acute Serous Otitis Media) is 15 and the predicted result from the network is also 15. All records are predicted correctly. Thus, the accuracy percentage for class A is 100%

- Horizontal reading from diagnostic results for ear disease class B (Cholesteatoma) is 11 and the predicted result from the network is also 11. All records are predicted correctly. Thus, the accuracy percentage for class B is 100%.

- Horizontal reading from diagnostic results for ear disease class C (Chronic Serous Otitis Media) is 43 and the predicted result from the network is also 43. All records are predicted correctly. Thus, the accuracy percentage for class C is 100%.

- Horizontal reading from diagnostic results for ear disease class D (Conductive Hearing Loss) is 10 and the predicted result from the network is also 10. All records are predicted correctly. Thus, the accuracy percentage for class D is 100%.

- Horizontal reading from diagnostic results for ear disease class E (Otosclerosis) is 17 and the predicted result from the network is also 17. All records are predicted correctly. Thus, the accuracy percentage for class E is 100%.

- Horizontal reading from diagnostic results for ear disease class F (Sensorineural Hearing Loss) is 7 and the predicted result from the network is also 7. All records are predicted correctly. Thus, the accuracy percentage for class F is 100%.

4.2.2 Training with Naïve Bayes with PCA

From the previous experiment using NN between topic 4.1.1 with data set without PCA and 4.2.1 with data set with PCA, both experimental results are the same even though the dimention is not the same. Now, Naïve bayes is used as training algorithm for the experiment.

		Prediction result					
		A	B	C	D	E	F
Diagnosis result	A	15	0	0	0	0	0
	B	0	11	0	0	0	0
	C	0	0	43	0	0	0
	D	0	0	0	10	0	0
	E	0	0	0	0	17	0
	F	0	0	0	0	0	7

Table 11: Confusion matrix of naïve bayes is trained on data set with PCA

Table 11 displays the results using NB on data set with PCA technique. The confusion matrix table can be interpreted as follows:

- Horizontal reading from diagnostic results for ear disease class A (Acute Serous Otitis Media) is 15 and the predicted result from the network is also 15. All records are predicted correctly. Thus, the accuracy percentage for class A is 100%

- Horizontal reading from diagnostic results for ear disease class B (Cholesteatoma) is 11 and the predicted result from the network is also 11. All records are predicted correctly. Thus, the accuracy percentage for class B is 100%.

- Horizontal reading from diagnostic results for ear disease class C (Chronic Serous Otitis Media) is 43 and the predicted result from the network is also 43. All records are predicted correctly. Thus, the accuracy percentage for class C is 100%.

- Horizontal reading from diagnostic results for ear disease class D (Conductive Hearing Loss) is 10 and the predicted result from the network is also 10. All records are predicted correctly. Thus, the accuracy percentage for class D is 100%.

- Horizontal reading from diagnostic results for ear disease class E (Otosclerosis) is 17 and the predicted result from the network is also 17. All records are predicted correctly. Thus, the accuracy percentage for class E is 100%.

- Horizontal reading from diagnostic results for ear disease class F (Sensorineural Hearing Loss) is 7 and the predicted result from the network is also 7. All records are predicted correctly. Thus, the accuracy percentage for class F is 100%.

4.2.3 Training with Decision tree with PCA

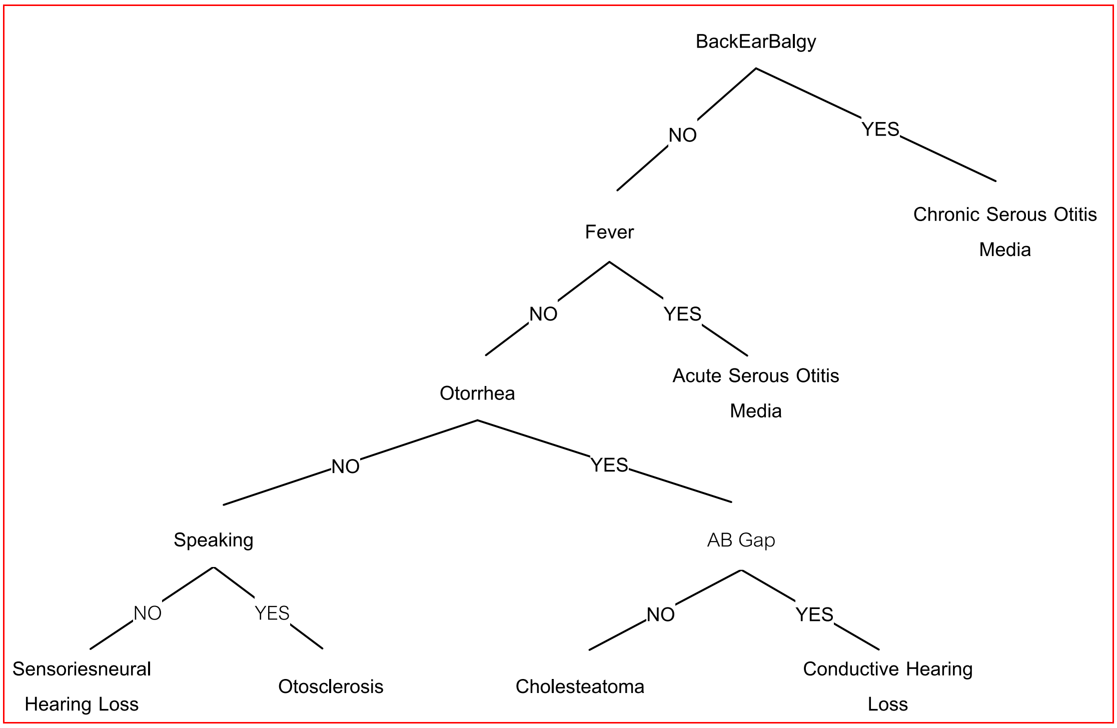


Figure 11: The result from training using decision tree on data set with PCA.

Figure 11 shows the result from the training using decision tree. We can go through the decision tree as the top-down approach until the leaf node is reached. For example, at the root node of the graph, if the BackEarBalgy attribute value is "YES", the predicting result is Chronic Serous Otitis Media class. However, if the value of the attribute BackEarBalgy is "NO", and the Fever attribute value is "NO". Then, the predicted class is Otorrhea. If attribute BackEarBalgy value is "NO", and the Fever attribute value is "YES", then the class is Acute Serous Otitis Media. The interpretation goes on in this manner until the leave nodes predicted class are reached.

The training using GINI split method can be discuss as follows:

Step one:

Attributes	GINI Split
BackEarBulgy	0.455197133
Confussion	0.455197133
RetractedEardrum	0.455197133
Fever	0.533892852
Vertigo	0.557665689
BalanceProblem	0.574809644
Fullness	0.581486552
Otorrhea	0.587476113
FacialPulsy	0.590890937
AB Gap	0.645252803
Ototoxicity	0.650015361
Hearing Loss	0.659508234
Otogia	0.659508234
Speaking	0.659508234

Table 12: Result of GINI split in step one

Table 12 shows the GINI value calculated for each attribute. The values are shown in ascending order (lowest to highest). Generally, the splitted attribute should be the attribute with the lowest GINI value. Therefore, attributes: BackEarBulgy, Confusion and RetractedEardrum can be choosen as the splitted attributes. Here, BackEarBulgy is selected.

Step two:

Attributes	GINI Split
Fever	0.541956882
BalanceProblem	0.564194458
Fullness	0.566113625
Vertigo	0.568773449
FacialPulsy	0.568773449
Otorrhea	0.585218703
AB Gap	0.617251462
Ototoxicity	0.627259259
Speaking	0.637891738
Hearing Loss	0.637891738
Otogia	0.637891738
RetractedEardrum	0.783950617
Confussion	0.783950617

Table 13 : Result of GINI split in step two

The coresponding GINI value for each attributed is recomputed. The results of GINI values for all attributes are shown in ascending order as Table 13. In generally, the splitted attribute is the one with the lowest GINI value. Fever is the attribute with the lowest value. Therefore, Fever attribute is chosen as the split attribute.

Step three:

Attributes	GINI Split
Otorrhea	0.46491865
Fullness	0.478782558
BalanceProblem	0.484108538
AB Gap	0.502393692
Ototoxicity	0.51637964
Speaking	0.528629579
Hearing Loss	0.528629579
Vertigo	0.528629579
Otogia	0.528629579
FacialPulsy	0.528629579
RetractedEardrum	0.728001782
Confussion	0.728001782

Table 14 : Result of GINI split in step three

Table 14 shows the GINI values that are re-calculated for all the corresponding attributes. From the results in table 14, Otorrhea gives the lowest GINI value. Therefore, this attribute is chosen as the split attribute.

Step four:

Attributes	GINI Split
Speaking	0
BalanceProblem	0
Hearing Loss	0
Vertigo	0
Otogia	0
FacialPulsy	0
Ototoxicity	0.412912913
Fullness	0.438276114
AB Gap	0.438276114
RetractedEardrum	0.438276114
Confussion	0.438276114

Table 15 : Result of GINI split in step four

Table 15 shows the GINI values that are re-computed for their corresponding attributes. It can be seen that 6 attributes: Speaking, BalanceProblem, Hearing loss, Vertigo, Otogia, and FacialPulsy gave the lowest GINI value. Here, speaking is chosen as the split attribute.

Step five:

Attributes	GINI Split
AB Gap	0
BalanceProblem	0
Ototoxicity	0
Fullness	0
Hearing Loss	0.497777778
Vertigo	0.497777778
Otogia	0.497777778
FacialPulsy	0.497777778
RetractedEardrum	0.497777778
Confussion	0.497777778

Table 16 : Result of GINI split in step five

Table 16 shows the GINI value for each attribute in ascending order. Here AB Gap attribute is chosen as the split attribute.

		Prediction result					
		A	B	C	D	E	F
Diagnosis result	A	15	0	0	0	0	0
	B	0	11	0	0	0	0
	C	0	0	43	0	0	0
	D	0	0	0	10	0	0
	E	0	0	0	0	17	0
	F	0	0	0	0	0	7

Table 17: Confusion matrix of decision tree is trained on data set with PCA

Table 17 illustrates the confusion matrix from decision tree. The result can be interpreted as follows:

The labels A, B, C, D, E and F denote ear disease classes. For example, A represents Acute Serous Otitis Media, B is Cholesteatoma, C is Chronic Serous Otitis Media, D is Conductive Hearing Loss, E is Otosclerosis and F is Sensorineural Hearing Loss. Performance of the Naïve bayes approach can be evaluated using numbers in the matrix. The examples of interpretations are:

- Horizontal reading from diagnostic results for ear disease class A (Acute Serous Otitis Media) is 15 and the predicted result from the network is also 15. All records are predicted correctly. Thus, the accuracy percentage of class A is 100%

- Horizontal reading from diagnostic results for ear disease class B (Cholesteatoma) is 11 and the predicted result from the network is also 11. All records are predicted correctly. Thus, the accuracy percentage of class B is 100%.

- Horizontal reading from diagnostic results for ear disease class C (Chronic Serous Otitis Media) is 43 and the predicted result from the network is also 43. All records are predicted correctly. Thus, the accuracy percentage of class C is 100%.

- Horizontal reading from diagnostic results for ear disease class D (Conductive Hearing Loss) is 10 and the predicted result from the network is also 10. All records are predicted correctly. Thus, the accuracy percentage of class D is 100%.

- Horizontal reading from diagnostic results for ear disease class E (Otosclerosis) is 17 and the predicted result from the network is also 17. All records are predicted correctly. Thus, the accuracy percentage of class E is 100%.

- Horizontal reading from diagnostic results for ear disease class F (Sensorineural Hearing Loss) is 7 and the predicted result from the network is also 7. All records are predicted correctly. Thus, the accuracy percentage of class F is 100%.

It can be seen that the network correctly classified 103 records from a total test data set 103 records. The accuracy percentage is 100 %.

4.3 Experiments with the Cross-Validation

Cross-Validation algorithm is a technique, in which the objective is to decrease the error from prediction by splitting the raw data set into k-fold. One of them is selected as the test set and others are the training set. Repeat the process by choosing different test set. The process is stop when we reach k test set.

The error from experiments are then summed and divided by numbers of folds. Here, k is equal to 10. Repeat the experiments for each learning algorithm. The accuracy of the learning algorithm will be obtained.

4.3.1 Training with Artificial neural network

Number of test set	Accuracy percentage
1st	100.00%
2nd	100.00%
3rd	100.00%
4th	100.00%
5th	100.00%
6th	100.00%
7th	100.00%
8th	100.00%
9th	100.00%
10th	100.00%
CV.	100.00%

Table 18: The result from training using Artificial Neural Network with Cross-Validation

Table 18 displays the accuracy percentage of training the neural network with cross-validation. The accuracy percentage is 100% for all simulations. Note that, the network architecture is 35-12-6. The learning rate and momentum are set to 0.1. The activation function is sigmoid function

4.3.2 Training with Naïve Bayes

Number of test set	Accuracy percentage
1st	100.00%
2nd	100.00%
3rd	100.00%
4th	100.00%
5th	100.00%
6th	100.00%
7th	100.00%
8th	100.00%
9th	100.00%
10th	100.00%
CV.	100.00%

Table 19: The result using Naïve bayes learning algorithm with Cross-Validation

In this section, the 10 folds cross-validation is applied with Naïve Bayes learning algorithm. Table 19 shows the accuracy percentage for each simulation. It can be seen that each simulation has reach 100% accuracy in the same manner as training with neural network.

4.3.3. Training using Decision tree

Number of test set	Accuracy percentage
1st	100.00%
2nd	100.00%
3rd	100.00%
4th	100.00%
5th	100.00%
6th	100.00%
7th	100.00%
8th	100.00%
9th	100.00%
10th	100.00%
CV.	100.00%

Table 20: The result from training using Decision Tree with Cross-Validation

Decision tree algorithm with 10 fold cross-validations is run. Table 20 shows the accuracy percentage for each simulation. The accuracy percentage for each simulation can reach 100%

CHAPTER 5
CONCLUSION DISCUSSION AND FUTURE WORK

5.1 Conclusion and Discussion

Diagnosis is important in medical process, the purpose is to classify the main symptom of diseases and suggest the correct treatment to the clinicians. The experimental results illustrate that techniques can be used this framework as a tool to assist the physicians in predicting the disease class from the basic signs and symptoms.

These three methodologies illustrate the potential for medical diagnosis. The result of classifying ear diseases is very encouraging with a hundred percent accuracy for all three techniques.

All patients in this study are severe cases so the clinician needs to provide the surgery treatment. Therefore, the diagnosis stages are very critical for the patients. The clinician expects the high accuracy result from this study. The experimental results in this study could be used as useful information for medical decision support.

Three main issues should be mentioned are:

- Comparing three input data sets
- Comparing three learning algorithms
- Commenting about Matlab

5.1.1 Comparing three input data sets (original data set, with PCA, with cross-validation)

The original data set consists of 34 input attributes but it can be effectively reduced to 14 input attributes using the principal component analysis. The experimental results are still 100% accuracy. Principal component analysis technique gives the advantage in this study. However, there is no additional advantage from cross-validation since the classification results have already reach 100% accuracy before applying the cross-validation technique.

5.1.2 Comparing three learning algorithms

Artificial neural networks have been succeeded in applying with various applications including classifying chief complaint in ear diseases [41]. However, training the network can be very slow. In addition, there are many parameters involved in the training such as momentum, learning rate, activation function and number of hidden nodes. Choosing the parameters which give the optimal results are tedious and time consuming.

For Naïve Bayes, the computations are much simpler than the neural networks. The training took less time than training the network. However, the accuracy performance is the same as the network in this study. Both algorithms results in 100% accuracy.

For decision tree, the algorithm is straight forward. To choose the split attribute, the GINI index for each attribute must be recomputed in each iteration. The result from training using decision tree algorithm is a tree-like graph. The graph shows the decision, their possible consequences and outcomes. Surprisingly, the tree-like graph that obtained from the experiment is the same as the treatment decision made by the IPD clinicians.

5.1.3 Commenting about Matlab

Matlab is a programming application for algorithm development, data analysis, image processing and numerical computation. The latest version (R20112a) of Matlab was selected as the experimental tool. This version has many wizards and tools to support users for instance nnstart, cvpartition and NaiveBayes.fit. The users should have some experiences in programming so that they can develop or adjust the desired algorithms. The coding is shown in appendix B. Results from Matlab can be displayed in the form of graph and image.

5.2 Future work

The medical diagnosis acts as a process of finding the possible disease. The clinicians try to use their knowledge and experience to determine the correct disease from signs and symptoms occurred. The hypothesis of disease is very critical mainly depending on the physician's analysis and vision. In some cases, the wrong diagnosis leads to the wrong treatments which are very risky for patients' health being or life.

This research is still at its beginning. The data set consists of only adult patients whose ages are over 15 years old. The classification of ear diseases in children patients are still remains. It is very difficult to get the correct symptoms from children because these young patients cannot explain or define their signs or symptoms clearly.

For future work, some other techniques will be explored and compared. It is interesting to see whether the accuracy percentage is still reach 100% when the number of class is increased.

REFERENCES

[1] Sofianita Mutalib, Nor Azlin Ali, Shuzlina Abdul Rahman and Azlinah Mohamed. An Exploratory Study in Classification Methods for Patients' Dataset. IEEE Conference on Data Mining and Optimization. (2009): 79-83.

[2] Konstantinos K. Delibasis, Pantelis A. Asvestas and George K. Matsopoulos. Computer-Aided Diagnosis of Thyroid Malignancy Using an Artificial Immune System Classification Algorithm. IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE. (2009): 680-686.

[3] Ning Yang-cui, Zheng Xiao-xian, Zhao Jing and Jiang Gui-juan. Forecasting the natural forest stand age based on artificial neural network model. CCTAE IEEE Conference Publications. (2010): 536-539.

[4] Rani P. Repeat Based Naive Bayes Classifier for Biological Sequences. ICDM IEEE Conference Publications. (2008): 989-994.

[5] Singh, M., Singh, P. and Hardeep Singh. Decision Tree Classifier for Human Protein Function Prediction. ADCOM IEEE Conference Publications. (2006): 564-568.

[6] Mark Hudson Beale, Martin T. Hagan and Howard B. Demuth. Neural network Toolbox User's Guide. The MathWorks inc. Online Only version.

[7] Clementine 12.0 User's Guide. http://www.spss.com. ISBN-13: 978-1-56827-395-2. Access on October 16, 2011.

[8] Fei fei Wang, Siripun Sanguasintukul and Chidchanok Lursinsap. Curl Forecasting for Paper Quality in Paper Making Industry. International Conference on Scientific Computing (ICSC 2008). China. (2008): 1281-1286.

[9] Naroumon Yordphet and Siripun Sanguansintukul. Safety Stock Based On Consumption Forecast by the Artificial Neural Network. The International Conference on Software and Computing Technology (ICSCT 2010). China.

[10] Yi Wang, Siripun Sanguasintukul and Chidchanok Lursinsap. The Customer Lifetime Value Prediction in Mobile Telecommunications. IEEE International Conference on Management of Innovation & Technology Publication. (2008).

[11] S. Areerachakul and S. Sanguansintukul. Clustering Analysis of Water Quality for Canals in Bangkok. The 2010 International Conference on Computational Science and ITS Applications. Eight IEEE/ACIS International; Conference on Computer and Information Science. (2009).

[12] J. Víctor Marcos, Roberto Hornero, Daniel Álvarez, Félix Del Campo and Miguel López. Applying Neural Network Classifiers in the Diagnosis of the Obstructive Sleep Apnea Syndrome from Nocturnal Pulse Oximetric Recordings. IEEE Computer Society (2007): 5174-5177.

[13] Paul S. Heckerling, Ben S. Gerber, Thomas G. Tape and Robert S. Wigton. Prediction of Community-Acquired Pneumonia Using Artificial Neural Networks. Medical Decision Making. (2003): 112-121.

[14] Ahmed M. Badawi, Manal Abdel Wahed, Shaimaa M. Elembaby. Diagnosis of Bladder Outlet Obstruction using Objective Parameters and Neural Networks Classifiers. (2004): 347-349.

[15] Wen-wei Ouyang, Xiao-zhong Lin, Yi Ren1 Yi Luo, Yun-tao Liu, Jia-min Yuan, Ai-hua Ou and Guo-zheng Li. TCM Syndromes Diagnostic Model of hypertension□study based on Tree Augmented Naive Bayes. (2011): 834-837.

[16] Susan P. Imberman Ph.D., Irene Ludwig, M.D. and Sarah Zelikovitz Ph.D. Using Decision Trees to Find Patterns in an Ophthalmology Dataset. (2011): 95:96.

[17] http://www.medscape.com. Access on November 27, 2010.

[18] Chai U-Sawat, Rada Dara, Rattinan Pirawanitkul and Siripun Sriwanyong. Anatomy & Physiology of the ear.

[19] http://www.emedicinehealth.com/script/main/art.asp?articlekey=138514&ref=129286. Access on November 27, 2010.

[20] Kenneth N. Anderson. MOSBY'S MEDICAL, NURSING, AND ALLIED HEALTH DICTIONARY. Fourth Edition.

[21] Theeraporn Ratana-anakchai and Supaporn Srirompothong. Otolaryngology Book, Khon-Kaen : Klangnanawittaya. (2008).

[22] L.Zhange, D. Wang and L.Chang. A Model on Forecasting Safety Stock of ERP based on BP neural Network. Proc.Management of Innovation and Technology. (2008): 1418-1422.

[23] http://www.codeproject.com/KB/recipes/NeuralNetwork_1.aspx. Access on October 15, 2011.

[24] http://www.willamette.edu/~gorr/classes/cs449/Classification/perceptron.html. Access on December 6, 2011.

[25] Frank Ronsenlatt. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. Psychological Review. (1958): 386-408.

[26] Marvin L. Minsky and Seymour A. Papert. Perceptrons: An Introduction to Computational Geometry. Cambridge MIT press. (1988): pp. 292.

[27] http://en.wikibooks.org/wiki/Artificial_Neural_Networks/Print_Version. Access on October 15, 2011.

[28] Pang-Ning Tan, Michael Steinbach and Vipin Kumar. General Approach to Solving a Classification Problem, Introduction to Data Mining. Boston : Pearson Addison-Wesley. (2005).

[29] http://www.sussex.ac.uk/Users/andrewop/Courses/NN/NNs5_6_MLP.ppt#309,63,Slide 63. Access on May 5, 2012.

[30] Rumelhart, D. E., Hinton, G. E. and Williams, R. J. Learning representations by back-propagating errors. (1986): 533–536.

[31] http://www.sussex.ac.uk/Users/andrewop/courses/NN/NNS5_6_MLP.ppt. Access on May 5, 2012.

[32] Ludmila I. Kuncheva. On the optimallity of Naive Bayes with dependent binary features. http://pages.bangor.ac.uk/~mas00a/papers/lkprl06.pdf. (2006): 830-837

[33] Gabriele Cevenini and Maria R. Massai. A naïve Bayes classifier for planning transfusion requirements in heart surgery. Journal of Evaluation in Clinical Practice. (2011): 1365-2753.

[34] http://www.mathworks.com/help/toolbox/stats/br039qw-1.html. Access on May 5, 2012.

[35] http://en.wikipedia.org/wiki/Naive_Bayes_classifier. Access on May 5, 2012.

[36] http://www.cse.unsw.edu.au/~billw/cs9414/notes/ml/06prop/id3/id3.html. Access on May 5, 2012.

[37] http://en.wikipedia.org/wiki/GINI_index. Access on May 5, 2012.

[38] http://en.wikipedia.org/wiki/Decision_tree_learning. Access on May 5, 2012.

[39] Karl Pearson. Principal Components Analysis. Edinburgh and Dublin Philosophical Magazine and Journal. (1901): 559-572.

[40] Douglas Trewartha and John Ebden. Investigating Data Mining in MATLAB. (2006).

[41] Ian H.Witten and Eibe Frank. Data Mining: Practical machine Learning Tools and Techniqueues. Morgan Kaufmann Publishers. 2nd edition. (2005).

[42] Robert Hecht-Nielsen. Theory of the Backpropagation Neural Network. International Conference (IJCNN). USA. (1989): I-593 – I-605.

[43] Narin Watanasusin and Siripun Sanguansintukul. Classifying Chief Complaint in Ear Diseases using Data Mining Techniques. Digital Content Multimedia Technology and its Application (IDCTA). (2011): 149-153.

[44] Budi Santosa. INTRODUCTION TO MATLAB NEURAL NETWORK TOOLBOX. (2002).

[45] Naïve bayes http://www.mathworks.com/help/toolbox/stats/naivebayes.fit.html. Access on February 10, 2012.

[46] http://www.mathworks.com/help/toolbox/stats/classregtreeclass.html, Decision Tree in MATLAB. Access on February 10, 2012.

APPENDICES

APPENDIX A

INTRODUCTION TO MATLAB

Introduction

Matlab is a tool for doing numerical computations with matrices and vectors. It can also display information graphically.

Definition of Variables

Variables are assigned numerical values by typing the expression directly, for example, typing:

```
a = 1+2
ans: a = 3
```

The answer will not be displayed when a semicolon is put at the end of an expression, for example type a = 1+2;

MATLAB utilizes the following arithmetic operators:

+	addition
-	subtraction
*	multiplication
/	division
^	power operator
'	transpose

A variable can be assigned using a formula that utilizes these operators and either numbers or previously defined variables. For example, since was defined previously, the following expression is valid:

```
b = 2*a;
```

To determine the value of a previously defined quantity, type the quantity by itself:

```
b
ans: b = 6
```

If your expression does not fit on one line, use an ellipsis (three or more periods at the end of the line) and continue on the next line.

```
c = 1+2+3+...
5+6+7;
```

There are several predefined variables which can be used at any time, in the same manner as user-defined variables:

i	sqrt(-1)
j	sqrt(-1)
pi	3.1416...

For example:

```
y= 2*(1+4*j)
ans: y = 2.0000 + 8.0000i
```

There are also a number of predefined functions that can be used when defining a variable. Some common functions that are:

abs	magnitude of a number (absolute value for real numbers)
angle	angle of a complex number, in radians
cos	cosine function, assumes argument is in radians
sin	sine function, assumes argument is in radians
exp	exponential function

For example, with y defined as above:

```
c = abs(y)
ans: c = 8.2462
c = angle(y)
ans: c = 1.3258
```

With a=3 as defined previously,

```
c = cos(a)
ans: c = -0.9900
c = exp(a)
ans: c = 20.0855
```

Note that exp can be used on complex numbers. For example, with y = 2+8i as defined above,

```
c = exp(y)
ans: c = -1.0751 + 7.3104i
```

which can be verified by using Euler's formula:

```
c = exp(2)cos(8) + je(exp)2sin(8)
```

Definition of Matrices

MATLAB is based on matrix and vector algebra; even scalars are treated as 1x1 matrices. Therefore, vector and matrix operations are as simple as common calculator operations.

Vectors can be defined in two ways. The first method is used for arbitrary elements:

```
v = [1 3 5 7];
```

creates a 1x4 vector with elements 1, 3, 5 and 7. Note that commas could have been used in place of spaces to separate the elements. Additional elements can be added to the vector:

```
v(5) = 8;
ans: v = [1 3 5 7 8].
```

Previously defined vectors can be used to define a new vector. For example, with v defined above

```
a = [9 10];
b = [v a];
```

creates the vector b = [1 3 5 7 8 9 10].

The second method is used for creating vectors with equally spaced elements:

```
t = 0:.1:10;
```

creates a 1x101 vector with the elements 0, .1, .2, .3,...,10. Note that the middle number defines the increment. If only two numbers are given, then the increment is set to a default of 1:

```
k = 0:10;
```

creates a 1x11 vector with the elements 0, 1, 2, ..., 10.

Matrices are defined by entering the elements row by row:

```
M = [1 2 4; 3 6 8];
```

creates the matrix

There are a number of special matrices that can be defined:

```
null matrix:  M = [ ];
nxm matrix of zeros:  M = zeros(n,m);
nxm matrix of ones:  M = ones(n,m);
nxn identity matrix:  M = eye(n);
```

A particular element of a matrix can be assigned:

```
M(1,2) = 5;
```

places the number 5 in the first row, second column.

Operations and functions that were defined for scalars in the previous section can also be used on vectors and matrices. For example,

```
a = [1 2 3];
b = [4 5 6];
c = a + b
ans:  c =  5  7   9
```

Functions are applied element by element. For example,

```
t = 0:10;
x = cos(2*t);
```

creates a vector x with elements equal to cos(2t) for t = 0, 1, 2, ..., 10.

Operations that need to be performed element-by-element can be accomplished by preceding the operation by a ".". For example, to obtain a vector x that contains the elements of x(t) = tcos(t) at specific points in time, you cannot simply multiply the vector t with the vector cos(t). Instead you multiply their elements together:

```
t = 0:10;
x = t.*cos(t);
```

APPENDIX B

CODING

I. Artificial Neural Network [42]

```
% Clear all parameters

clear;

clc;


% attribute

a = xlsread('ND_ANN.xls', 1, 'A2:AH258');


% class

c = xlsread('ND_ANN.xls', 1, 'AI2:AI258');

c_newff = xlsread('ND_ANN.xls', 1, 'AJ2:AO258');


% partition training and test set

holdout = cvpartition(c,'holdout',0.6); % 60 percent


% training and test attribute

training_attribute = a(holdout.training,:)';

test_attribute = a(holdout.test,:)';


% training and test class

training_class = c_newff(holdout.training,:)';

test_class = c_newff(holdout.test,:)';


% set inputs and training

inputs = training_attribute;

targets = training_class;
```

```
% set hidden node
hidden_node = 12;

% set epochs
num_epochs = 10000;

% set minimum error
min_error = 0.001;

% set learning rate
learning_rate = 0.1;

% set momentum
momentum = 0.1;

net = newff(inputs, targets, hidden_node, {'tansig', 'tansig'}, 'traingdm');

% set training
net.divideParam.trainRatio = 1;
net.divideParam.valRatio = 0;
net.divideParam.testRatio = 0;
net.trainParam.epochs = num_epochs;
net.trainParam.min_grad = min_error;
net.trainParam.lr = learning_rate;
net.trainParam.mc = momentum;

[net, tr] = train(net,test_attribute,test_class);
outputs = net(inputs);
[err, cm] = confusion(targets, outputs)
```

II. Naïve bayes [43]

```
% clear all parameters
clear;
clc;

% attribute
a = xlsread('ND.xls', 1, 'A2:AH258');

% class
[num, txt, raw] = xlsread('ND.xls', 1, 'AI2:AI258');
c = txt;

% partition training and test set
holdout = cvpartition(c,'holdout',0.4); % 40 percent

% training and test attribute
training_attribute = a(holdout.training,:);
test_attribute = a(holdout.test,:);

% training and test class
training_class = c(holdout.training,:);
test_class = c(holdout.test,:);

%nb = NaiveBayes.fit(training_attribute, training_class, 'Distribution', 'mvmn');
nb = NaiveBayes.fit(training_attribute, training_class, 'Distribution', 'kernel');

% predict
nb_predict = nb.predict(test_attribute);

% confusion matrix
nb_cmatrix = confusionmat(test_class, nb_predict)
```

III. Decision Tree [44]

```
% Clear all parameters
clear;
clc;

% attribute
a = xlsread('ND.xls', 1, 'A2:AH258');

% class
[num, txt, raw] = xlsread('ND.xls', 1, 'AI2:AI258');
c = txt;

% partition training and test set
holdout = cvpartition(c,'holdout',0.4); % 40 percent

% training and test attribute
training_attribute = a(holdout.training,:);
test_attribute = a(holdout.test,:);

% training and test class
training_class = c(holdout.training,:);
test_class = c(holdout.test,:);

% show tree Field name
t = classregtree(training_attribute, training_class);
view(t);

% add field name
names_value = {'Sex', 'Age', 'Otorrhea', 'Speaking', 'Tinnitus', ...
'AB Gap', 'Hearing Loss', 'Ototoxicity', 'NoiseInduce', 'EndocrineInduce', ...
'Articulatory Defect', 'Vertigo', 'FluidFillLevelHigh', 'Otogia', 'FacialPulsy', ...
'KeratinMass', 'EarOdor', 'Fullness', 'Trauma', 'FindingTerm', ...
'Heredity', 'DeafandPregnancy', 'Infection', 'Rhinitis', 'Fever', ...
'RedTempanicMembrane', 'RuptureEardrum', 'Bulging', 'BalanceProblem', …
'Dizziness', 'Headache', 'BackEarBulgy', 'RetractedEardrum', 'Confussion'};
t = classregtree(training_attribute, training_class, 'names', names_value);
view(t);

% predict
predict = eval(t, test_attribute);

% confusion matrix
t_cmatrix = confusionmat(test_class, predict)
```

APPENDIX C

Principal Component Analysis

In this research, we run the PCA by SPSS Clementine version 12.0 which it is the latest version. We will show how, as the following;

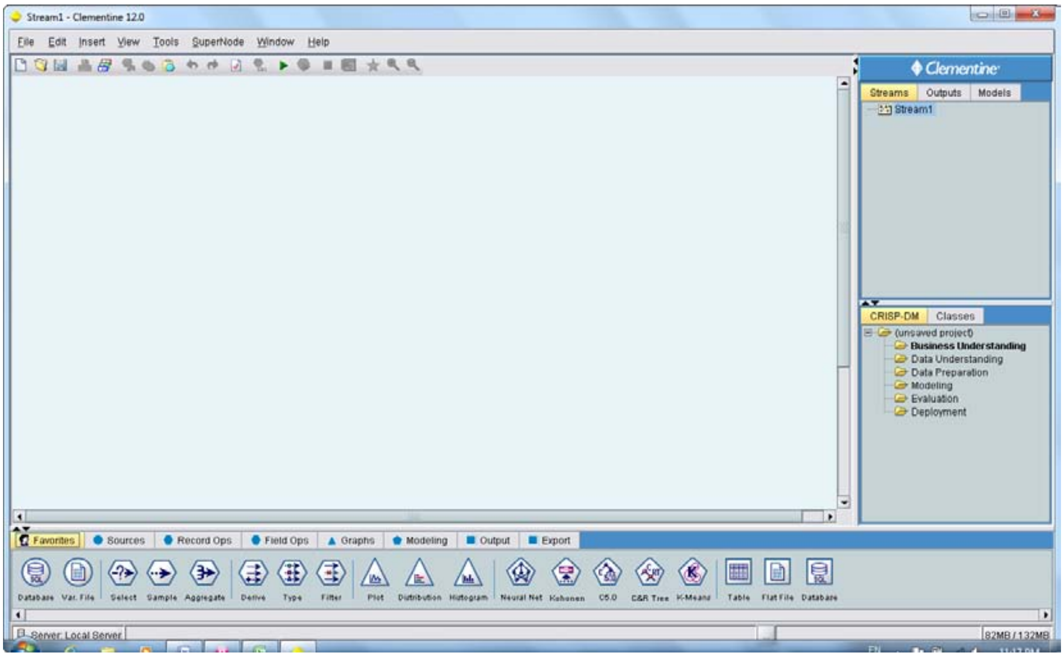

Stream1 - Clementine 12.0
File Edit Insert View Tools SuperNode Window Help
Clementine
Streams Outputs Models
Stream1
CRISP-DM Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites Sources Record Ops Field Ops Graphs Modeling Output Export
Database Var. File Select Sample Aggregate Derive Type Filter Plot Distribution Histogram Neural Net Kohonen C5.0 C&R Tree K-Means Table Flat File Database
Server: Local Server
82MB / 132MB

Figure 12: Clementine main application window

To start the application, choose Clementine 12.0 from the SPSS Inc program group on the Windows Start menu.

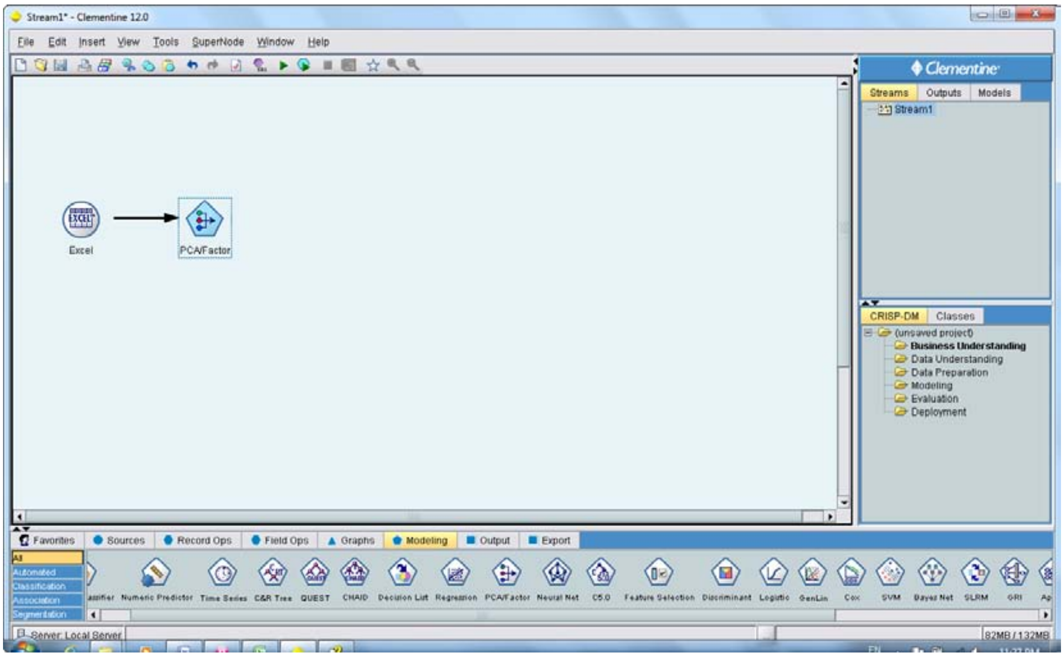
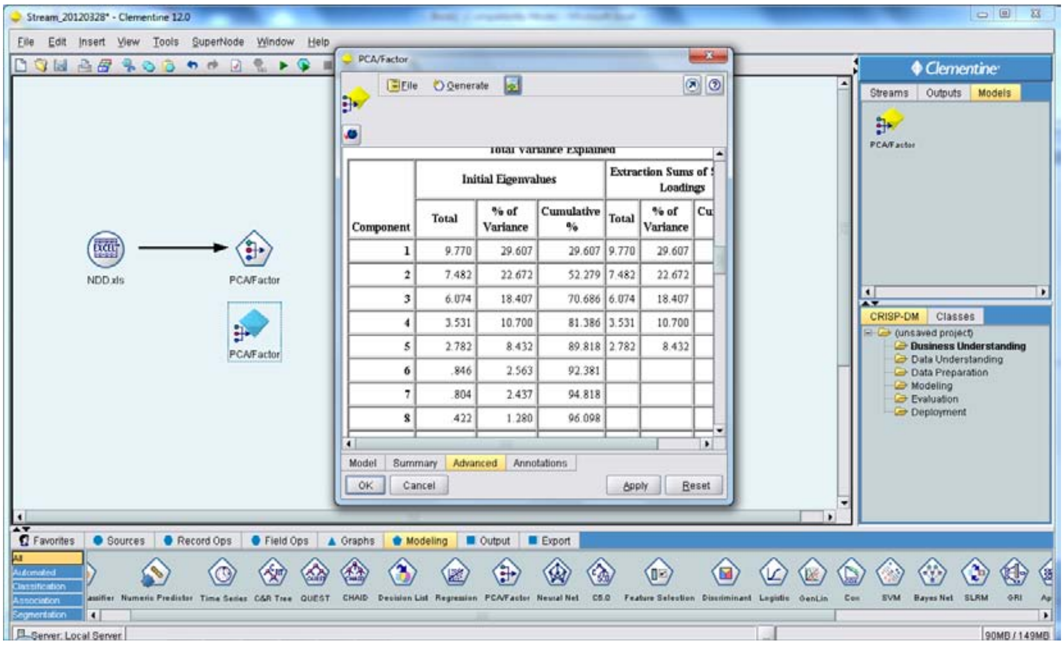


Figure 13: set the parameter

From figure 13, we select type of file from Sources tab then select PCA Factor from modeling tab.



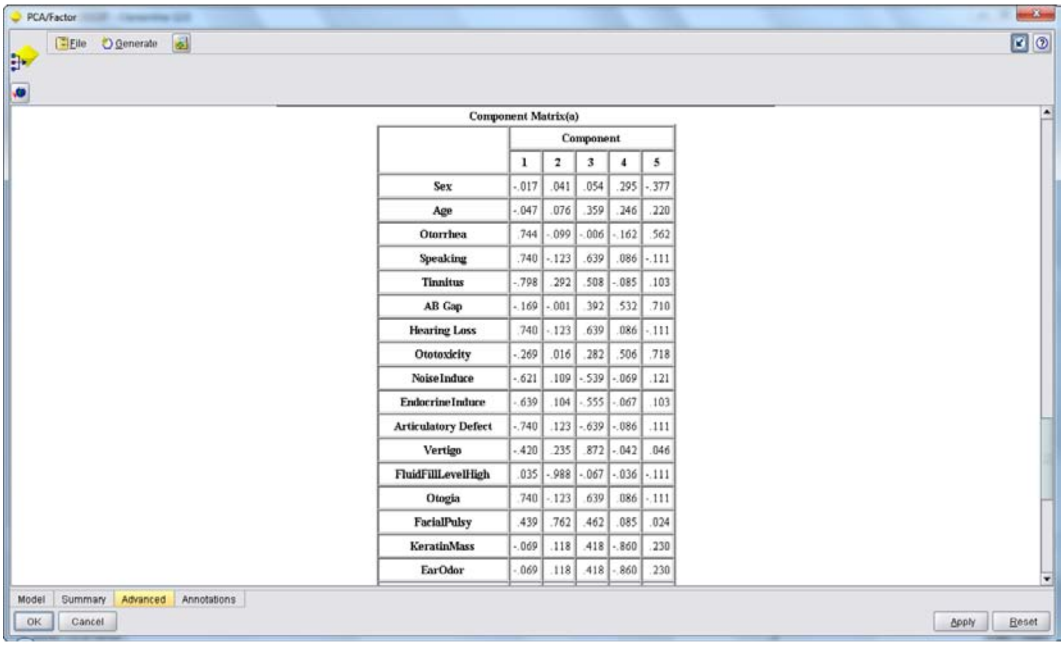
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of S Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cu
1	9.770	29.607	29.607	9.770	29.607	
2	7.482	22.672	52.279	7.482	22.672	
3	6.074	18.407	70.686	6.074	18.407	
4	3.531	10.700	81.386	3.531	10.700	
5	2.782	8.432	89.818	2.782	8.432	
6	.846	2.563	92.381			
7	.804	2.437	94.818			
8	.422	1.280	96.098			

Figure 14: result from PCA

From figure 14, we run the PCA then the result will show in the right hand side. If we want to see the result, double click on the icon PCA Factor. The window PCA Factor will show on screen.

On the PCA Factor table, we will get 5 components which get the accumulative percentage is 89.828%.

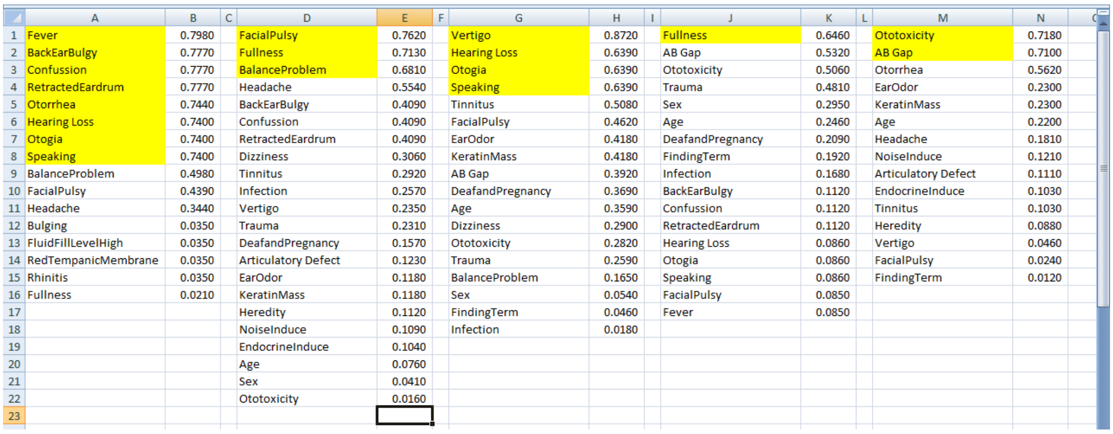


Component Matrix(a)

	Component				
	1	2	3	4	5
Sex	-.017	.041	.054	.295	-.377
Age	-.047	.076	.359	.246	.220
Otorrhea	.744	-.099	-.006	-.162	.562
Speaking	.740	-.123	.639	.086	-.111
Tinnitus	-.798	.292	.508	-.085	.103
AB Gap	-.169	-.001	.392	.532	.710
Hearing Loss	.740	-.123	.639	.086	-.111
Ototoxicity	-.269	.016	.282	.506	.718
NoiseInduce	-.621	.109	-.539	-.069	.121
EndocrineInduce	-.639	.104	-.555	-.067	.103
Articulatory Defect	-.740	.123	-.639	-.086	.111
Vertigo	-.420	.235	.872	-.042	.046
FluidFillLevelHigh	.035	-.988	-.067	-.036	-.111
Otogia	.740	-.123	.639	.086	-.111
FacialPulsy	.439	.762	.462	.085	.024
KeratinMass	-.069	.118	.418	-.860	.230
EarOdor	-.069	.118	.418	-.860	.230

Figure 15: detail of 5 components

From the figure 15, if we scroll down from figure15, we will see the detail of those components which the first column is the attribute name and component column is the value from PCA algorithm.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Fever	0.7980		FacialPulsy	0.7620		Vertigo	0.8720		Fullness	0.6460		Ototoxicity	0.7180
2	BackEarBulgy	0.7770		Fullness	0.7130		Hearing Loss	0.6390		AB Gap	0.5320		AB Gap	0.7100
3	Confussion	0.7770		BalanceProblem	0.6810		Otogia	0.6390		Ototoxicity	0.5060		Otorrhea	0.5620
4	RetractedEardrum	0.7770		Headache	0.5540		Speaking	0.6390		Trauma	0.4810		EarOdor	0.2300
5	Otorrhea	0.7440		BackEarBulgy	0.4090		Tinnitus	0.5080		Sex	0.2950		KeratinMass	0.2300
6	Hearing Loss	0.7400		Confussion	0.4090		FacialPulsy	0.4620		Age	0.2460		Age	0.2200
7	Otogia	0.7400		RetractedEardrum	0.4090		EarOdor	0.4180		DeafandPregnancy	0.2090		Headache	0.1810
8	Speaking	0.7400		Dizziness	0.3060		KeratinMass	0.4180		FindingTerm	0.1920		NoiseInduce	0.1210
9	BalanceProblem	0.4980		Tinnitus	0.2920		AB Gap	0.3920		Infection	0.1680		Articulatory Defect	0.1110
10	FacialPulsy	0.4390		Infection	0.2570		DeafandPregnancy	0.3690		BackEarBulgy	0.1120		EndocrineInduce	0.1030
11	Headache	0.3440		Vertigo	0.2350		Age	0.3590		Confussion	0.1120		Tinnitus	0.1030
12	Bulging	0.0350		Trauma	0.2310		Dizziness	0.2900		RetractedEardrum	0.1120		Heredity	0.0880
13	FluidFillLevelHigh	0.0350		DeafandPregnancy	0.1570		Ototoxicity	0.2820		Hearing Loss	0.0860		Vertigo	0.0460
14	RedTempanicMembrane	0.0350		Articulatory Defect	0.1230		Trauma	0.2590		Otogia	0.0860		FacialPulsy	0.0240
15	Rhinitis	0.0350		EarOdor	0.1180		BalanceProblem	0.1650		Speaking	0.0860		FindingTerm	0.0120
16	Fullness	0.0210		KeratinMass	0.1180		Sex	0.0540		FacialPulsy	0.0850			
17				Heredity	0.1120		FindingTerm	0.0460		Fever	0.0850			
18				NoiseInduce	0.1090		Infection	0.0180						
19				EndocrineInduce	0.1040									
20				Age	0.0760									
21				Sex	0.0410									
22				Ototoxicity	0.0160									
23														

Figure 16: analyze the 5 components

From the figure 16, we export the detail of 5 components to spreadsheet. We sort by descending. We consider only the positive result and select the higher value. In this cast, we select the value more over 0.6 which 14 attributes were selected.

VITAE

Narin Watanasusin was born in September 27th, 1975, in Bangkok, Thailand. He obtained his Bachelor's degree in Costing Accounting from the Account Faculty, The University of the Thai Chamber of Commerce in 1998. He has worked for Benchachinda Holding Company limited since graduated.