ANALYSIS OF MEDICAL DATA VIA ARTIFICIAL NEURAL NETWORKS WITH COMPARING DIFFERENT TRAINING METHODS AND ACTIVATION FUNCTIONS

by

Mehmet Fatih USLU

June 2011

ANALYSIS OF MEDICAL DATA VIA ARTIFICIAL NEURAL NETWORKS WITH COMPARING DIFFERENT TRAINING METHODS AND ACTIVATION FUNCTIONS

by

Mehmet Fatih USLU

A thesis submitted to

the Graduate Institute of Sciences and Engineering

of

Fatih University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Engineering

June 2011
Istanbul, Turkey

APPROVAL PAGE

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

_____
Assoc. Prof. Veli HAKKOYMAZ

This is to certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

_____
Dr. İhsan Haluk AKIN
Supervisor

Examining Committee Members

Assist. Prof. Tuğrul YANIK _____

Assist. Prof. Mustafa PETEK _____

Dr. İhsan Haluk AKIN _____

It is approved that this thesis has been written in compliance with the formatting rules laid down by the Graduate Institute of Sciences and Engineering.

_____
Assoc. Prof. Nurullah ARSLAN
Director

June 2011

ANALYSIS OF MEDICAL DATA VIA ARTIFICIAL NEURAL NETWORKS WITH COMPARING DIFFERENT TRAINING METHODS AND ACTIVATION FUNCTIONS

Mehmet Fatih USLU

M. S. Thesis - Computer Engineering
June 2011

Supervisor: Dr. İhsan Haluk AKIN

ABSTRACT

This study is about analyzing medical data with using neural networks and the accuracy of these networks on diagnosis. During the training of neural networks, a program named FannTool that is under GPL license and uses an artificial neural network library named Fast Artificial Neural Network is used. Training with different algorithms, different network designs, using data sets with different training/test ratio, training with different number of epochs, using different activation functions are investigated to find out how they affect the performance of artificial neural networks.

Keywords: Artificial Neural Networks, Anemia, Complete Blood Count.

YAPAY SİNİR AĞLARI İLE TIBBİ VERİLERİN DEĞİŞİK EĞİTİM METODLARI VE AKTİVASYON FONKSİYONLARI KULLANILARAK ANALİZİ

Mehmet Fatih USLU

Yüksek Lisans Tezi – Bilgisayar Mühendisliği
Haziran 2011

Tez Danışmanı: Öğr. Gör. Dr. İhsan Haluk AKIN

ÖZ

Bu çalışma tıbbi verilerin yapay sinir ağları yolu ile analizi ve bu ağların tanı koyabilmedeki performansı ile ilgilidir. Ağlar eğitilirken Fast Artificial Neural Network isimli bir yapay sinir ağı kütüphanesini kullanan FannTool isimli ücretsiz dağıtım bir program kullanılmıştır. Farklı algoritmalarla eğitilmenin, farklı ağ dizaynlarının, farklı eğitim/test oranına sahip veri setlerinin kullanımının, farklı miktarlarda adımlarla eğitilmenin, farklı aktivasyon fonksiyonları kullanımının yapay sinir ağının performansını nasıl etkilediği araştırılmıştır.

Anahtar Kelimeler: Yapay Sinir Ağları, Anemi, Tam Kan Sayımı.

To my parents

ACKNOWLEDGEMENT

I express my appreciation to Assist. Prof. İhsan Ömür Bucak from Mevlana University and my brother Süleyman Uslu from Bosphorus University Computer Engineering Department for their support throughout the research.

I also express my appreciation to Mustafa Gül from Istanbul University Istanbul Faculty of Medicine, Mehmet Hamza Türkcanoğlu from Istanbul University Cerrahpaşa Faculty of Medicine and Ahmet Faruk Feriz from Hacettepe University Faculty of Medicine for their support on diagnosing procedure of Anemia.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

LIST OF SYSMBOLS AND ABBREVIATIONS

SYMBOL/ABBREVIATION

ANN	Artificial Neural Network
BASO	Basophil
BASO#	Amount of Basophil in the blood
BASO%	Percentage of Basophil in the blood
CBC	Complete Blood Count
EO	Eosinophil
EO#	Amount of Eosinophil in the blood
EO%	Percentage of Eosinophil in the blood
FANN	Fast Artificial Neural Network
fL	femto Liter
g/dL	gram per deci Liter
Gb/s	Gigabits per second
HCT	Hematocrit
HGB	Hemoglobin
LYMPH	Lymphocyte
LYMPH#	Amount of Lymphocyte in the blood
LYMPH%	Percentage of Lymphocyte in the blood
MCH	Mean Corpuscular Hemoglobin
MCHC	Mean Corpuscular Hemoglobin Concentration
MCV	Mean Corpuscular Volume
mmol/L	millimoles per Liter
MONO	Monocyte

MONO#	Amount of Monocyte in the blood
MONO%	Percentage of Monocyte in the blood
MPV	Mean Platelet Volume
MSE	Mean Squared Error
NEU	Neutrophil
NEU#	Amount of Neutrophils in the blood.
NEUT%	Percentage of Neutrophils in the blood.
PCT	Plateletcrit
PDW	Platelet Distribution Width
pg	pico Gram
PLT	Platelet Count
RBC	Red Blood Cell Count
RDW-CV	RBC Distribution Width
µL	micro Liter
WBC	White Blood Cell Count
Yr	Year

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Decision-making is one of the most important missions in medical science. Doctors are always required to make the right decisions. Despite the study and experience of tens of years, they might make mistakes as human beings (Sprogar, et al., 2002). Identifying complicated correlations which are hard to be revealed by human without the help of computation power of computers is succeeded with pattern recognition methods in which Artificial Neural Network (ANN) model is involved (Valafar and Ersoy, 1996). Huge training is required for artificial neural networks to find a pattern among a pattern pool each of which look like similar. Revealing the particular properties that artificial neural networks use is problematic.Existence or vacancy of a property makes ANNs to determine classes (Valafar, et al., 2000). As a result, in the last decades, making decisions according to computer-aided programs become more of an issue and the researches in this subject are extended (Allahverdi, et al., 2011). In this study, we tried to diagnose Anemia according to blood sample reports. For this, we collect blood samples of healthy people and people with Anemia. What we want to experience is that we use these blood sample reports and train an ANN with these samples. Then we test the trained ANN with the remainder samples and try to find medically meaningful results from those blood samples. In this study, by medically meaningful will refer to two different cases. First one is determining whether a person is

an Anemia patient or he or she is healthy. Second one has four choices which are healthy or three different Anemia types; microcytic, normocytic and macrocytic.

Firstly, in this introduction part, we investigated the similar researches in this field and saw what they have done. Then we thought what we could do further and talk about them. In chapter 2, we speak of the neural networks, structure and kinds of ANNs. In chapter 3, we give some medical background information that is significant for understanding this paper. In chapter 4, we give the core of our study and its results in detail. And in conclusion part, we discuss the results and talk about the future study.

We gained the most successful results from FANN_TRAIL_INCREMENTAL method. In binary choice test, we get 98.40% success, which means we successfully find whether a person has Anemia or not. In 4-ary test, which we consider the type of illness, we get 95.80% success.

1.2 PRELIMINARY RESEARCH

Our aim in this study is to help doctors in diagnosis of Anemia. We have used Fast ANN, which is a kind of ANNs under GPL.

There are similar studies that have been done in this area that is the assistance of computer-aided programs in medical science. We have examined several researches related to Anemia diagnosis or treatment. In addition, in these articles, there are several methods besides ANNs have been used and we had a chance to check the comparison of the results of different techniques.

The first paper was about therapy of Anemia patients with chronic renal failure. They have studied for a better treatment of that specific group of Anemia patients due to the cost of the process and to reduce the side effects by individualizing the drug dosage. 110 patients have been examined and several ANN models used including MLP, Elman and FIR. In all methods, time-series process is exercised. The results were satisfactory. Thus, an easy-to-use decision-aid computer program is implemented (Guerrero, et al., 2003).

Another article was written on the subject that is the effect of Hydroxyurea (affected or not affected) on sickle Anemia patients. The effects are examined on 83 patients over 23 parameters. The first method used was correlation analysis and it is failed in finding a relation between Hemoglobin and these 23 parameters. The second method was linear regression and it is also failed. The successful method is an ANN which uses pattern recognition analysis and it is successful with a rate of 86.6% on the effect of Hydroxyurea on the examined patients (Valafar, et al., 2000).

The subject of the last study was very close to our paper. It indicates that in the Hematology Labs, the number of Full Blood Counts (FBC), their accuracy and the number of parameters in FBCs have increased. This leads a raw data to be examined by Hematology experts. The idea is that an algorithm can be used in order to conclude with medically meaningful data from those raw data generated by FBCs. Bayer R&D team writes an ANN with a hidden layer between input and output nodes. Blood of 1000 patients from 22 hematology centers are examined. ANN which is trained by labeled samples gives successful results. Moreover, it is stated that in the future, not only binary results (patient or not patient) but also the type of the disease can be achieved using neural networks (Zini and d'Onofrio, 2003).

Looking at the studies on medical assistance of computer programs, even if the other methods do not give satisfactory results, ANNs end up with significantly better models. The results tell us that we gain about 90-95% success during this Neural Network study.

The important part of the study is that the articles which we have examined are all about determining if the patient has that illness or not. They have successfully diagnosed the patients. Key point is that they answered binary questions like if the patient is affected or he is healthy. We, here, also try to decide if someone has Anemia depending on the blood test results. Afterwards, we want to find what kind of cell leads people to Anemia. At this point, we have 4 choices, he is healthy (1), or he has Anemia with microcytic (2), normocytic (3) or macrocytic (4) cells.

CHAPTER 2

NEURAL NETWORKS

2.1 GENERAL

Neural networks or ANNs were seen to be the next generation of the programming architectures. They do not need a programmer for learning, they can learn by themselves. This new kind of programming technique was started to be used by many of the computer scientist with a big hope. However, there was a disillusion for those who rely on this new programming technique. There were some of those who tried but failed in neural networks. These neural networks were complicated and confusing. There are only a few neural network architectures currently commercially available. Many of those are in research. They are being worked on in laboratories all over the world (Anderson and McNeill, 1992).

The reason why the scientists are concentrated on neural networks that much is that they observe animals and see that they have a very fast decision mechanism compared to machines. Classical methods on observing objects and deciding movements according to these objects, for instance, require so much computational power. An idea to overcome this problem is to come up with a new programming architecture similar to those in animals' brains as the scientist believe and this new type of programming isANNs (Mehrotra, et al., 1996).

One of the first studies about neural networks can be considered as the study of Warren McCulloh, neurophysiologist, and Walter Pitts, mathematician, in 1943. They wrote a paper about how the neurons might work in human brain and developed a model for electrical circuits. In 1949, Donald Hebb wrote a book named 'Organization of Behavior'. He claimed that the paths between neurons get stronger as they are used (Anderson and McNeill, 1992).

After 1950, with the developments in computer science area, it is proposed that neural networks can be used in programming. One of the first scientists tried to use neural networks in programming was Nathanian Rochester in IBM Research Labs. In 1956, artificial intelligence and neural networks used together in Darthmouth Summer Research Project and in both areas there has been improvements. After this project, John von Neuman and Frank Rosenblatt do some research on neural networks in computer science. In 1985 and 1987, American Institute of Physics and Institute of Electrical and Electronic Engineering arrange meetings about neural networks (Mehrotra, et al., 1996).

Today, neural networks are studied in many areas and these studies are in a key position for technology.

2.2 METHODS

It cannot be considered as true to say that there is only one type of neural networks. Different kinds of neural network architectures have been proposed for different type of problems. Same architectural concept is also in the structure of human brain. For instance, the cerebral cortex, where is thought to do the most of the processing in human brain, consists of 5 to 7 layers each of which connected by taking inputs from and giving outputs to other layers (Izhikevich, 2006).

Back-Propagation Algorithm: This algorithm started to be developed by different researchers in different locations in early 1970s. Currently it is the most common, easy to learn algorithm and applicable to multi-layered and complicated structures. Its main advantage is to propose nonlinear solutions to ill-defined problems. Usually there is an

input and an output layer with a hidden layer. Number of hidden layers may change but it is thought that three hidden layer are enough to solve a problem with any complexity (IBM Research, 2005).

Other uses of neural networks are below:

Feedforward Neural Networks: The first and the simplest neural networks. Information flows to forward. There is no cycle or loop (Anderson and McNeill, 1992).

Stochastic Neural Networks: Adds random variants to the networks (Anderson and McNeill, 1992).

Physical Neural Networks: This type of neural networks are used in electrical circuits. Resistances of resistors are adjustable in these types of circuits (Anderson and McNeill, 1992).

Spiking Neural Networks: These types of neural networks take into consideration the timings of inputs. The advantage is that it can process the information in time domain, which means processing of time varying signals is possible in spiking neural networks (Anderson and McNeill, 1992).

Dynamic Neural Networks: Deals with nonlinear behaviors with multi variables and behaviors that are independent of time (Anderson and McNeill, 1992).

Cascading Neural Networks: Starts with a tiny network and then trains itself, adds new nodes and becomes a multi-layered network (Anderson and McNeill, 1992).

2.3 APPLICATION AREAS

Animal behaviors that do not need mechanical movement such as perception and recognition require neural networks. The idea of applying these behaviors to machines means the start of ANNs. General types of these behaviors are classification, clustering, vector quantization, pattern association, function approximation, forecasting, control application, optimization and search. Some of these are described below:

In classification, new observed objects are classified according to previous classes that are determined before. For this process, it is needed to have a data set that consists of classified objects. By inspecting on this data set, new objects can be classified.

Forecasting is a wide-used application type from stock markets to weather casting. Actually forecasting is similar to function approximating in high level. Feedforward and recurring networks are important for and used in forecasting (Dematos, 1996)

Optimization is very significant for business areas. The process is about optimizing, that is minimizing or maximizing, with respect to specific constraints. For instance, minimizing the cost or printing area of a circuit board designed for a specific purpose. Stochastic networks are used in optimization area (Mehrotra, et al., 1996).

CHAPTER 3

MEDICAL INFORMATION

3.1 COMPLETE BLOOD COUNT

Complete Blood Count (CBC) is a kind of medical tests. It includes basically hemoglobin, hematocrit, mean cell hemoglobin, mean cell hemoglobin concentration, mean cell volume, platelet count, red blood cell count and red cell distribution width. It can include additional parameters on these.

BASO: (Basophil) When you apply basic dye to those white blood cells that have coarse granules on them, they become blue. 1 percent or less of white blood cells is basophils and it is subjected to change in some conditions like diseases. (AACC, 2005)

BASO#: Amount of basophil in the blood

BASO%: Percentage of basophil in the blood

EO: (Eosinophil) They are produced in bone marrow. Their number in body is between 0 and 450. They generate 1-3% of white blood cells (Collins, 2004).

EO#: Amount of eosinophil in the blood

EO%: Percentage of eosinophil in the blood

HCT: (Hematocrit) The ratio of the volume of the red blood cells to the total volume of the blood (Purves, et al., 2004).

HGB: (Hemoglobin) It is a protein that carries oxygen from blood to tissues and carbon monoxide from tissues to blood (Anthea, 1993).

LYMPH: (Lymphocyte) A kind of white blood cell and it consists almost half of them (Dorlands, 2007).

LYMPH#: Amount of lymphocyte in the blood.

LYMPH%:percentage oflymphocyte in the blood sample.

MCH: (Mean Corpuscular Hemoglobin) Average amount of hemoglobin per red blood cell in the blood (MedlinePlus, 2009).

MCHC: (Mean Corpuscular Hemoglobin Concentration) Concentration of hemoglobin of sample red blood cells that is between 4.9 to 5.5 mmol/L (Van Beekvelt, et al., 2001).

MCV: (Mean Corpuscular Volume) Average red blood cell volume. Reference is between 80-100 fL (MedlinePlus, 2009).

MONO: (Monocyte) Kind of white blood cell, half of which are stored in spleen (Swirski FK, et al., 2009)

MONO#: Amount of monocyte in the blood

MONO%: Percentage of monocyte in the blood

MPV: (Mean Platelet Volume) Average size of platelets in the blood. Younger platelets are larger than older ones.

NEU: (Neutrophil) Type of white blood cell either segmented or banded. Generates PMNs with basophils and eosinophils (Witko-Sarsat, V, 2000).

NEU#: Amount of neutrophils in the blood.

NEUT%: Percentage of neutrophils in the blood.

PCT: (Plateletcrit) Ratio of the total platelets to the blood (Wiwanitkit V., 2004)

PDW: (Platelet Distribution Width) Platelet size distribution. It indicates active platelet release.

PLT: (Platelet Count) Amount of platelets in the blood.

RBC: (Red Blood Cell Count) Amount of red blood cells in the blood.

RDW-CV: (RBC Distribution Width) Distribution of red blood cell count.

WBC: (White Blood Cell Count) Amount of white blood cells in the blood.

Some elements distribution on CBC graphics can be seen in Figure 3.1, in Figure 3.2 and in Figure 3.3. These data is taken from a National Health Nutrition Examination Survey (NHANES) research result. Columns represent CBC element amount and rows represent how many people's blood has element at this interval. For example in the Figure 3.1, about 1600 people's blood has between 4.296 and 4.528 red blood cells in a micro liter.

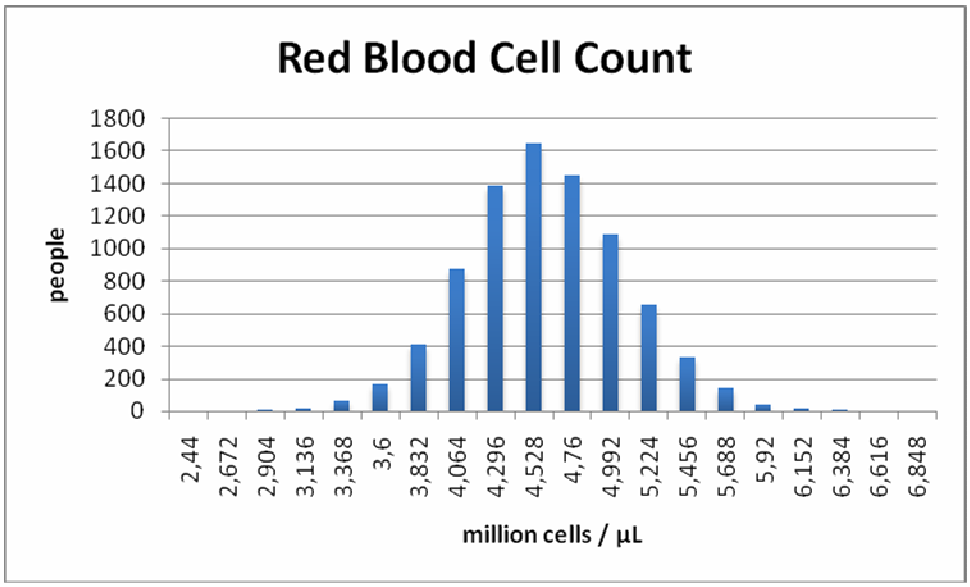

Figure 3.1 Normal Distribution of Red Blood Cell Count (NHANES, 2006)

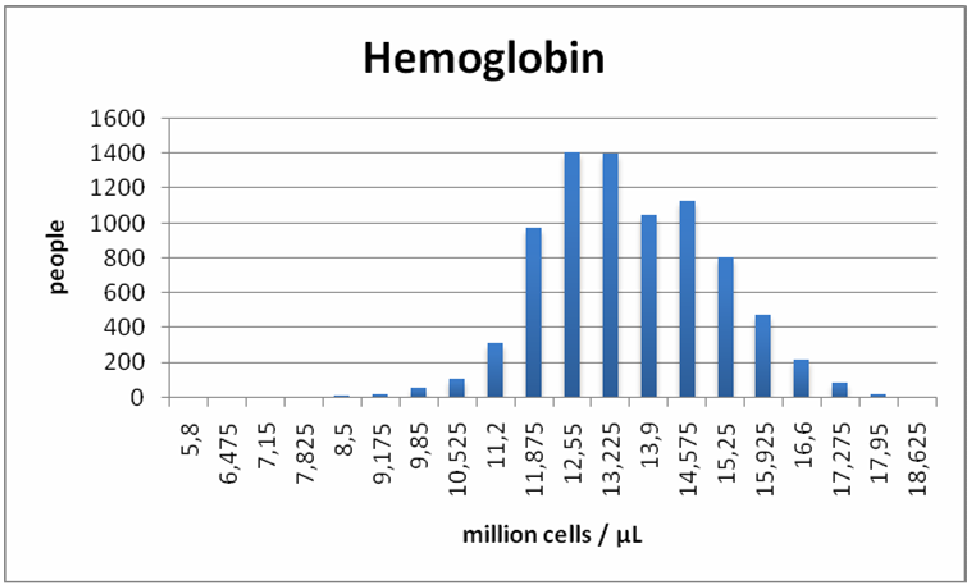

Figure 3.2 Normal Distribution of Hemoglobin (NHANES, 2006)

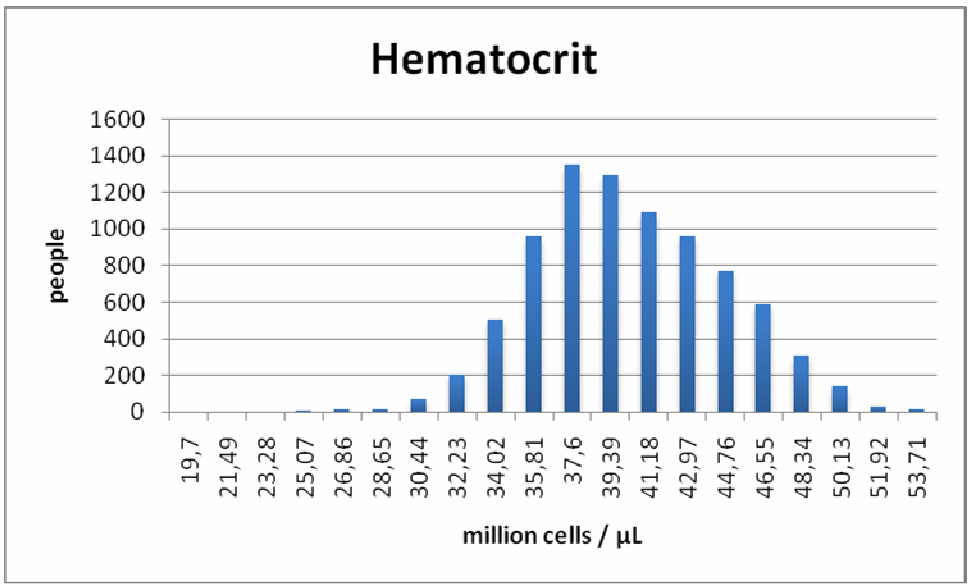

Figure 3.3 Normal Distribution of Hematocrit (NHANES, 2006)

3.2 ANEMIA

Anemia, which is known as bloodlessness, can be caused by many factors. The symptoms are usually tiredness, weakness, dizziness, tinnitus, spots on eyes, pallor and headache. The first test that should be done on a suspicious Anemia case is hemogram test.

International Statistical Classification of Diseases and Related Health Problems (ICD) codes are the worldwide acceptable disease classification system. Anemia is between D50-D64 in ICD-10 codes which indicates that there are many kinds of Anemia. The purpose of the project is to figure out if the patient has Anemia. ANN is a technology which is developed by the inspiration of the data processing mechanism of human brain. By using ANN, simple biologic neural systems can be simulated. ANNs produce solutions for problems which require human observation and thought. In this project, doctors' capability of diagnosis of different kinds of Anemia using the results and symptoms of CBC will be tested. (WHO, 2007)

CBC (Complete blood count) or FBC/FBE (Full blood count/exam) is a test which gives an idea about the blood cell count, which is known as hemogram, of the patient. It can be used for many diseases but it is very significant information for Anemia diagnosis. Project is about interpreting the blood test results in terms of Anemia. The purpose is not making a diagnosis but helping the doctor in diagnosing process (Eisenstaedt R, et al., 2006).

3.2.1 Diagnosis

Anemia is basically diagnosed on CBC with hemoglobin, hematocrit and red blood cell count. These parameters' normal value range is in Table 3.1. Its data are taken from lecture notes of Istanbul University Cerrahpaşa Faculty of Medicine.

Table 3.1 Normal and Lower Limit values of Hemoglobin, Hematocrit and MCV (Dallman, 1977)

	Hemoglobin (g/dl)		Hematocrit (%)		Mean Corpuscular Volume (fL)	
Age (Year)	**Mean**	**Lower Limit**	**Mean**	**Lower Limit**	**Mean**	**Lower Limit**
0.5-1.9	12.5	11.0	37	33	77	70
2-4	12.5	11.0	38	34	79	73
5-7	13.0	11.5	39	35	81	75
8-11	13.5	12.0	40	36	83	76
12-14						
Female	13.5	12.0	41	36	85	78
Male	14.0	12.5	43	37	84	77
15-17						
Female	14.0	12.0	41	36	87	79
Male	15.0	13.0	46	38	86	78
18-49						
Female	14.0	12.0	42	37	90	80
Male	16.0	14.0	47	40	90	80

Red Blue Cell's lower limit value is 4.0 on females and 4.5 on males. Anemia can be diagnosed via observing hemoglobin, hematocrit or RBC decrease. In this thesis, samples include 18-59 years old patients CBCs, so hemoglobin lower limit is 16.0 on males and 14.0 on females. Some other sources give lower limit of hemoglobin 14.0 on males and 12.0 on females.

An Anemia can be microcytic, normocytic or macrocytic. Size of the erythrocytes (Mean Corpuscular Value) is decisive on diagnose of Anemia. If MCV is smaller than 80, we can say Anemia type is presumably microcytic. Iron Deficiency Anemia, Thalassemia or some other chronic disorders can cause microcytic Anemia. And if MCV is bigger than 100, this time Anemia type is likely macrocytic. Vitamin B12 lack or folic acid lack can cause macrocytic Anemia. If MCV is between 80 and 100, we can say there is maybe normocytic Anemia (Dallman, 1977).

CHAPTER 4

EXPERIMENTAL STUDY

4.1 DATA SOURCE

In the thesis, the main purpose was trying to simulate diagnosing process of the doctors on a specific area like CBC (Complete Blood Count). So, main requirement was CBC Results.

Obtaining real data was a complicated process in the research. For getting medical data from a university to another one, a lot of bureaucratic process is needed. In this research, firstly Fatih University gave a report from Scientific-Ethics Committee for requesting data from Istanbul University Istanbul Faculty of Medicine. Then, with this report, required CBC data could be taken from laboratory. Privacy of patients was an important issue; therefore patient identity information is not taken.

Number of Data: We have total 1688 CBC result at first. 1347 CBC are from male patients, 341 are from females. But some patients CBC is corrupted, so these CBCs eliminated. After this elimination, data set consists of 1606 CBC, 1278 male and 328 female. Youngest person is 18 years old, oldest is 49 and born year distribution of samples can be seen in Figure 4.1. In the Figure 4.2, Figure 4.3 and Figure 4.4 HCT, HGB and RBC distribution of patients graphic can be seen. Columns represent CBC elements amount and rows represent how many samples' values are at this level. For

example in the Figure 4.2, we can say about 400 patients' HCT value of blood are between 20 and 25 million cells / μL.

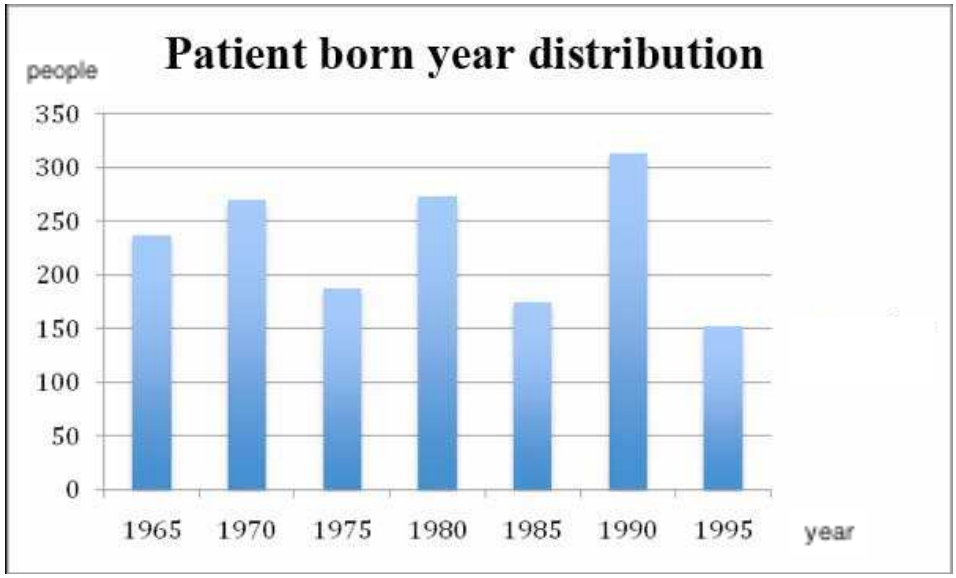

Figure 4.1 Born year distribution of samples

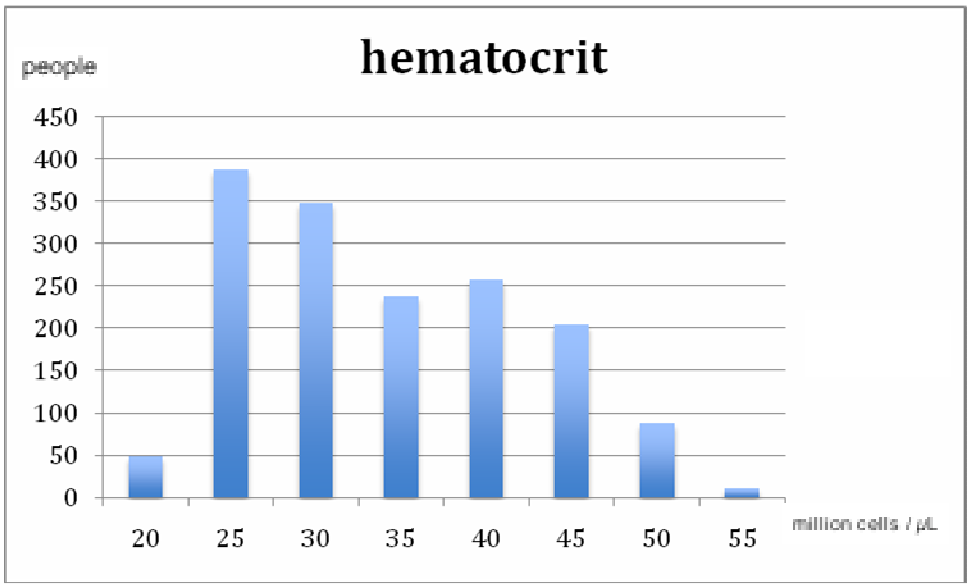

Figure 4.2 Hematocrit value distribution of samples

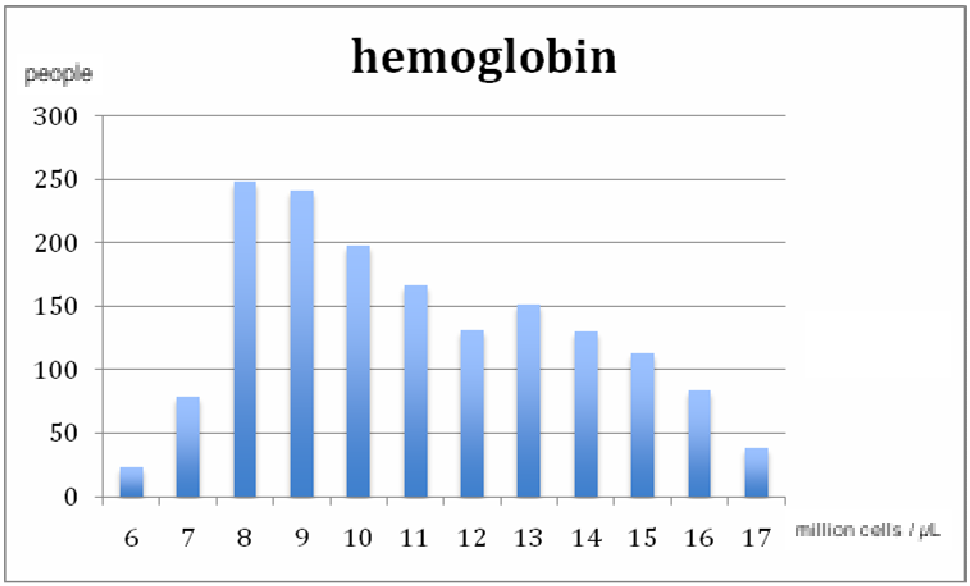

Figure 4.3 Hemoglobin value distribution of samples

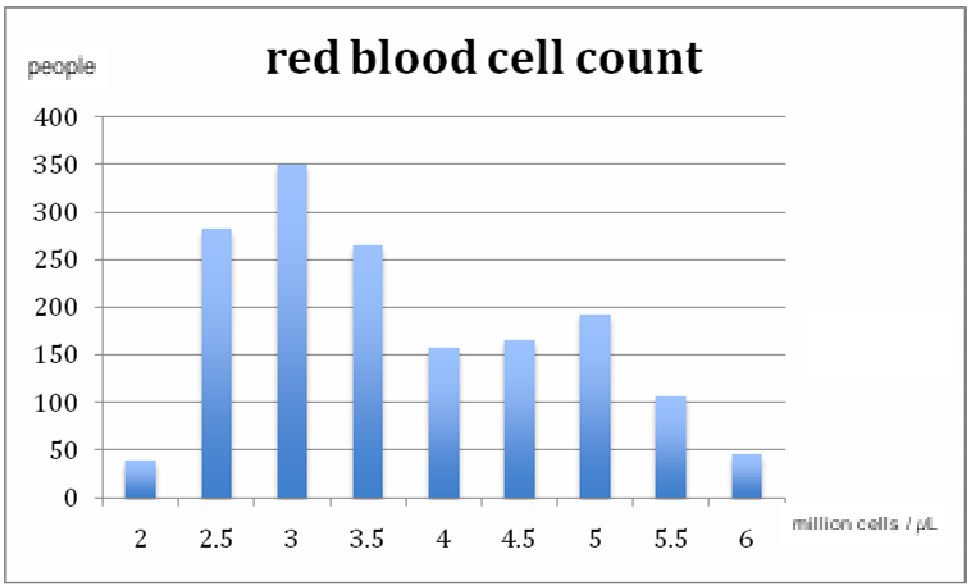

Figure 4.4 RBC value distribution of samples

4.2 FEATURE REDUCTION

Feature extraction can be considered as reducing the dimensions. Some of CBC elements are not effective on Anemia diagnosis. *ttest* function of MATLAB is used for this purpose.

4.2.1 TTest

Some of CBC elements and correlations on Anemia diagnosis graphics are below. A dot represents a sample, blue dots represent males and red dots represent females.Row groups represent diagnosis of this sample and columns represent CBC element value of sample. In Figure 4.5 correlation of hematocrit concentration and Anemia type of our samples can be seen. Correlation of hemoglobin concentration and Anemia type of our samples can be seen in Figure 4.6. Correlation of RBC and Anemia type of our samples can be seen in Figure 4.8. These three parameter are efficient for diagnosing Anemia but not efficient subtype of Anemia. In the Figure 4.8, mean cell value distribution and Anemia diagnosis graphic can be seen. MCV value are related Anemia type but not related diagnosis of Anemia. In the Figure 4.9, BASO concentration and Anemia relation graphic can be seen. BASO is a sample CBC element of not relative with Anemia. Therefore, ttest is applied to understand which CBC elements are efficient to diagnose Anemia and which are not.

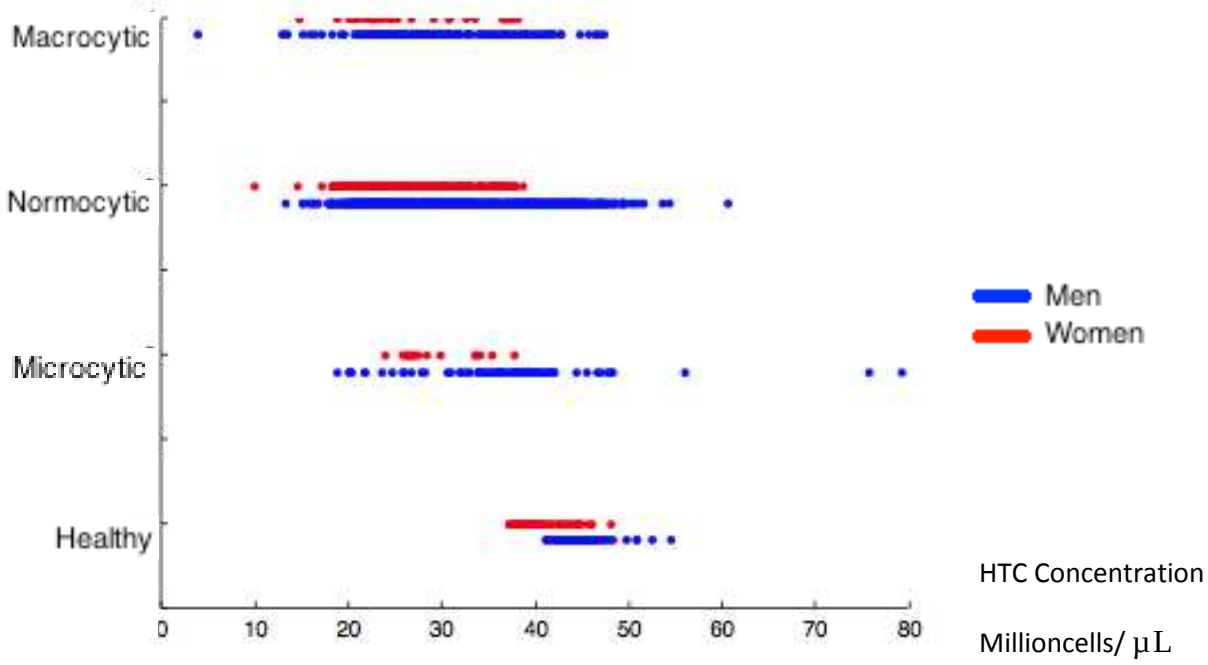

Figure 4.5 HCT and Anemia correlation

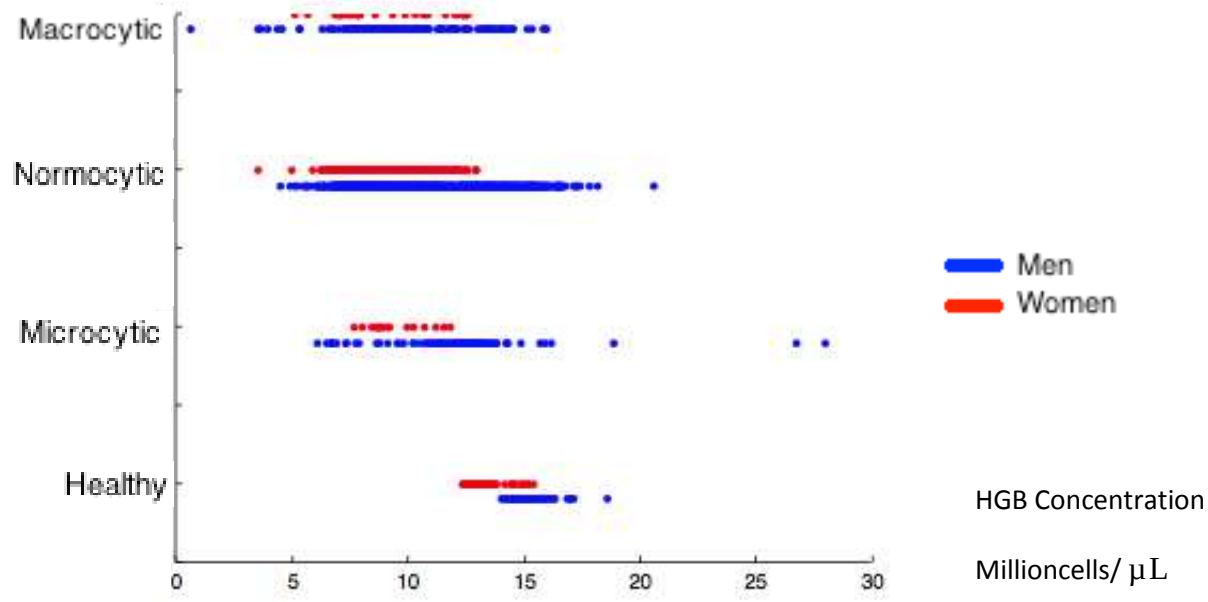

Figure 4.6 HGB and Anemia correlation

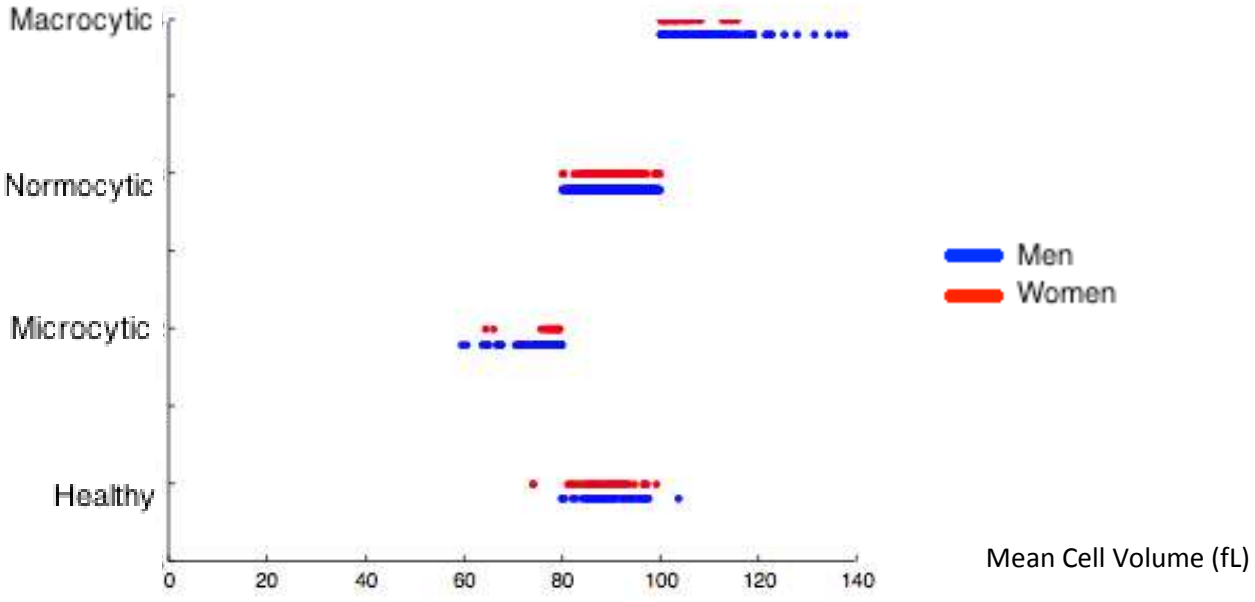

Figure 4.7 MCV and Anemia correlation

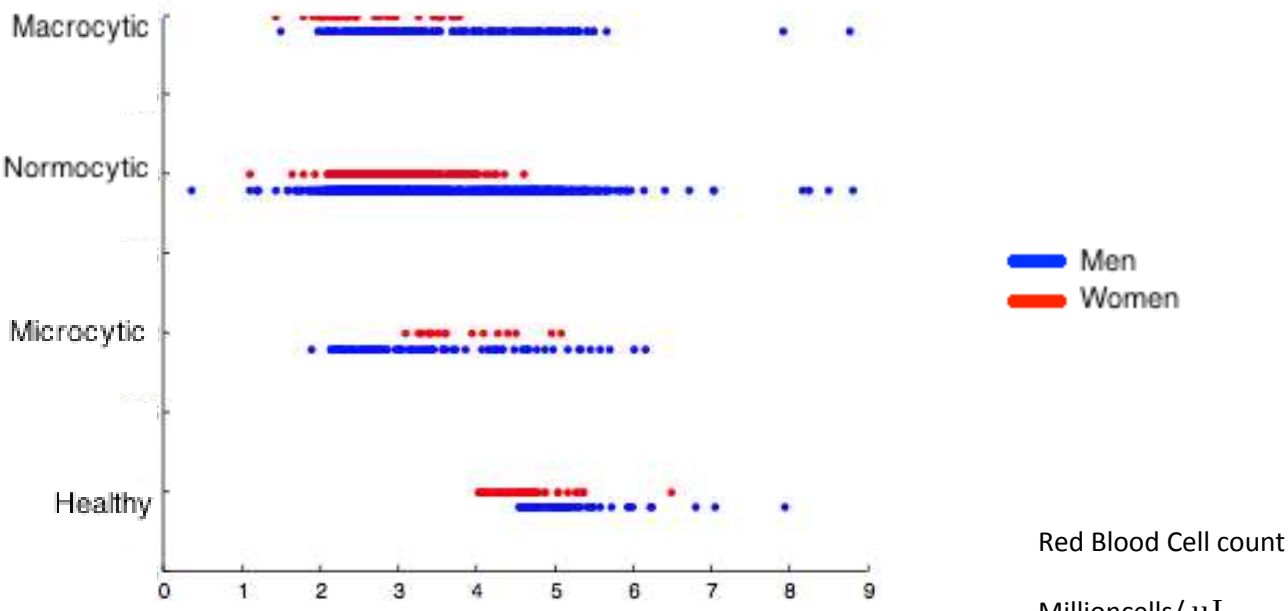

Figure 4.8 RBC and Anemia correlation

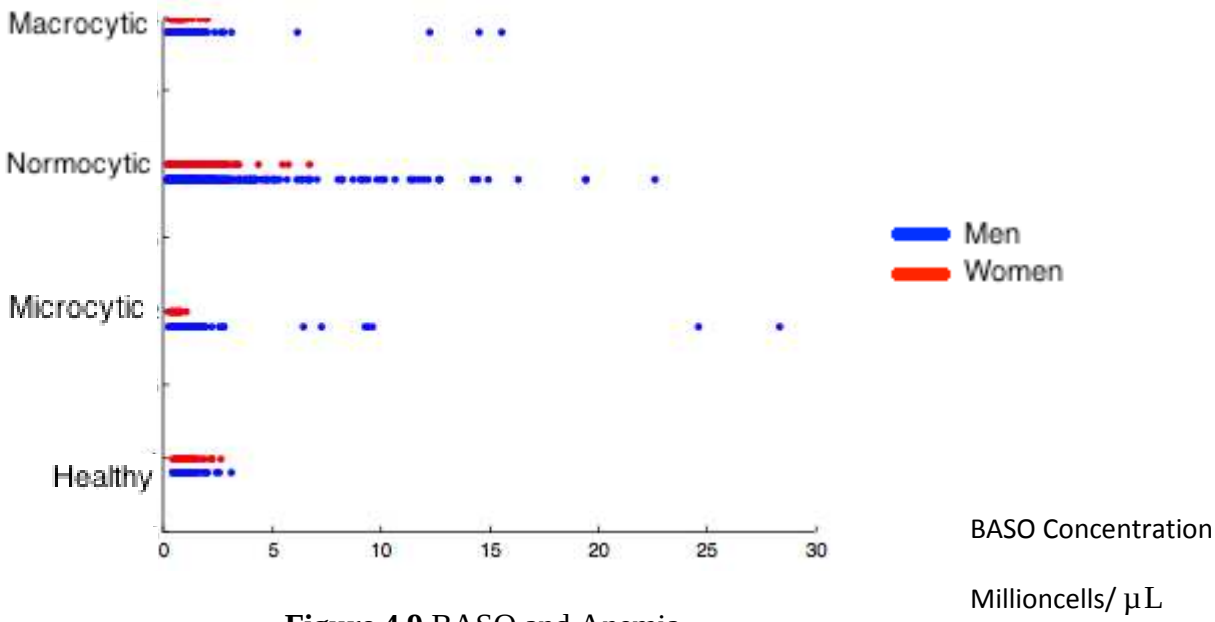

Figure 4.9 BASO and Anemia

CBC element's *ttest* results are in **Table 4.1**. Elements that have gray background color are successful in test, so it means they may be related by Anemia status.

Table 4.1 TTest results

Order	Title	Ttest result
1	Gender	1
2	Birth year	0
3	BASO#	0
4	BASO%	0
5	EO#	0
6	EO%	0
7	HCT	1
8	HGB	1
9	LYMPH#	1
10	LYMPH%	0
11	MCH	1
12	MCHC	1
13	MCV	1
14	MONO#	0
15	MONO%	0
16	MPV	1
17	NEU#	0
18	NEUT%	0
19	PCT	1
20	PDW	0
21	PLT	1
22	RBC	1
23	RDW-CV	0
24	WBC	1

4.2.2 Algorithm Comparison

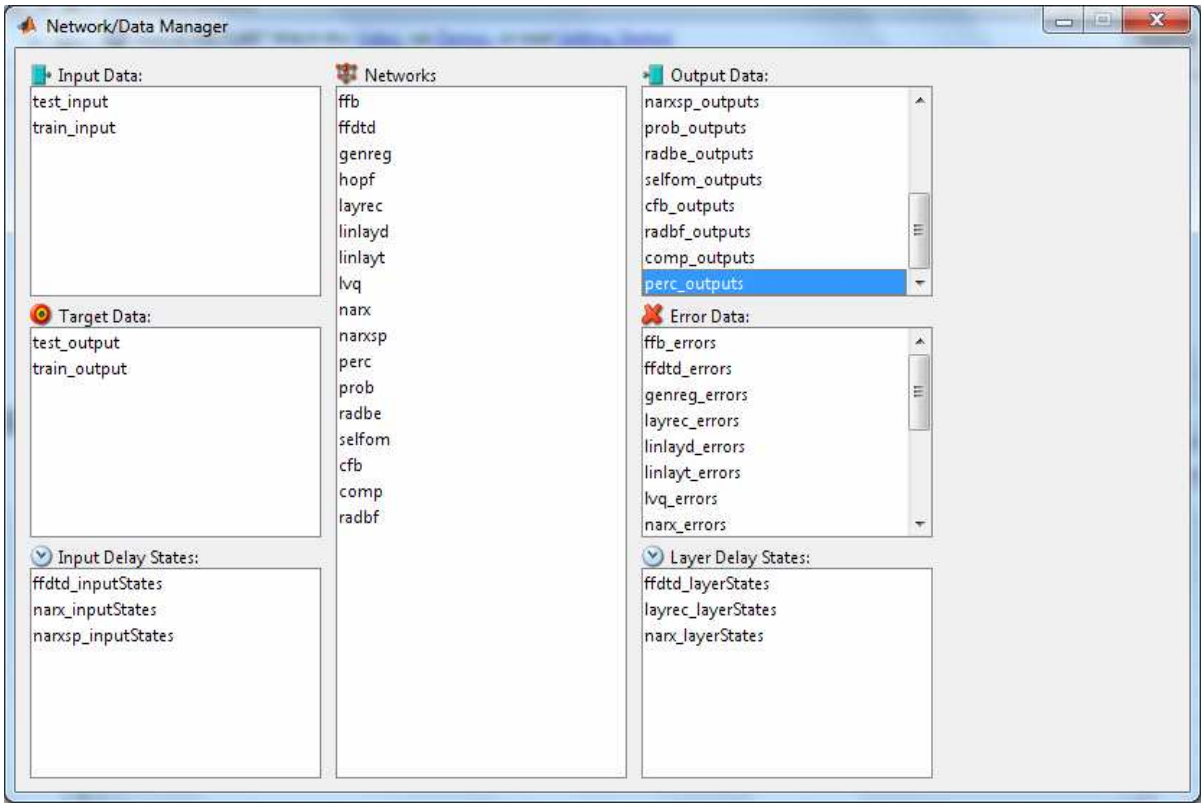

Figure 4.10 MATLAB neural networks screenshot

There are about 15 neural network train algorithms in MATLAB tool as seen in Figure 4.10. These are:

- Feed-Forward Backpropagation (ffb)
- Feed-Forward Distributed Time Delay (ffdtd)
- Generalized Regression (genreg)
- Hopfield (hopf)
- Layer Recurrent (layrec)
- Linear Layer (design) (linlayd)
- Linear Layer (train) (linlayt)
- LVQ (lvq)
- NARX (narx)
- NARX Series-Parallel (narxsp)
- Perceptron (perc)
- Probabilistic (prob)
- Radial Basis (exact fit) (radbe)
- Self-Organizing Map (selfom)
- Cascade-Forward Backpropagation (cfb)
- Competitive (comp)

- Radial Basis (fewer neurons) (radbf)

Performances of neural networks which use these algorithms are below. We showed the ones with performance higher than %90 in gray background in the Table 4.2.

Table 4.2 Comparing different ANN's performance.

ANN type	True	Total	Success
ffb	701	803	87.3%
ffdtd	723	803	90.0%
genreg	601	803	74.8%
hopf		803	0
layrec	727	803	90.5%
linlayd	721	803	89.8%
linlayt		803	0
lvq		803	0
narx	635	803	79.1%
narxsp		803	0
perc	105	803	13.1%
prob	105	803	13.1%
radbe	695	803	86.6%
selfom		803	0
cfb	754	803	93.9%
comp	105	803	13.1%
radbf	695	803	86.6%

4.3 TRAINING METHOD COMPARISON TOOL

Fast Artificial Neural Network Library (FANN) is a Multi-Layer Feed Forward Back Propagate ANN library written in C. It can be used in a lot of programming language like Perl, PHP, Java, and Delphi or C#. It has LGPL (GNU Library General Public License) license.

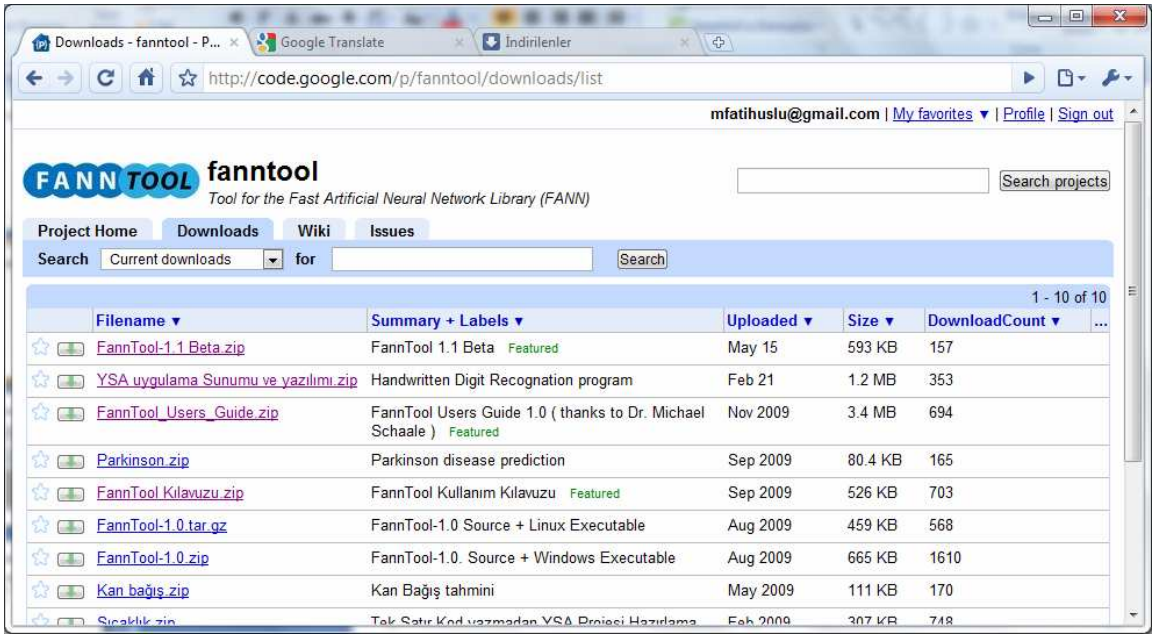

Figure 4.11 Screenshot of FannTool Page in Google Code

4.3.1 FANNTOOL MANUAL

FannTool is a program, which enables graphical user interface for FANN Library. In this research FannTool is used for ANN library. Program interface can be seen in Figure 4.12 and screenshot of FannTool page in Google Code web site can be seen in Figure 4.11.

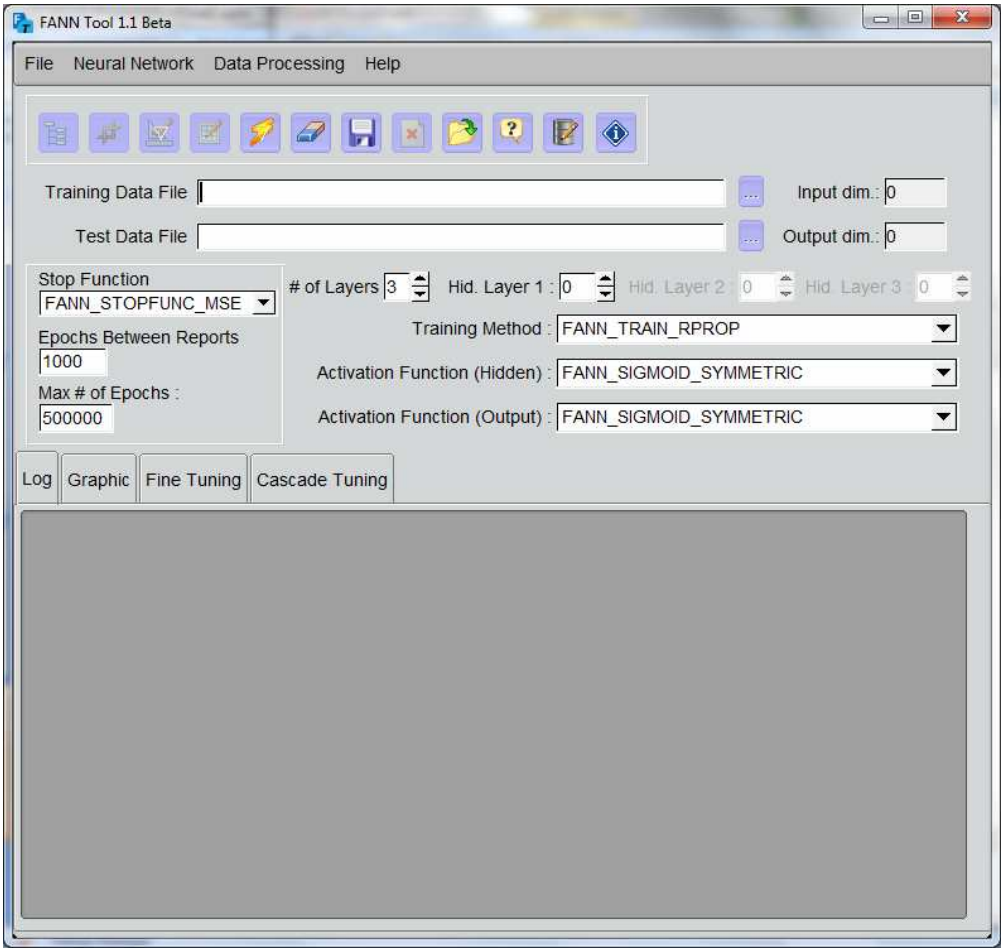

Figure 4.12 FannTool Program Interface

This is the main screen of FannTool. With this tool, firstly we have to make a Raw Data file. Then choose Number of Input and Number of Output. There is an option for Scale data and shuffle data. The Raw Data matrix must be in the format shown in Table 4.3.

Table 4.3 Data Raw Matrix

Inputs(x)	Outputs(y)	Number of data (n)
i11 i12 …	i1x o11 o12 …	o1y
i21 i22 …	i2x o21 o22 …	o2y
…		
in1 in2 …	inx on1 on2 …	ony

In “Data Processing” menu, you can choose raw data file and make some adjustments like training-testing fraction of data, number of input and output, shuffle and scale data.

Number of Input: In a sample row of our matrix, it shows amount of input values.

Number of Output: In a sample row of our matrix, it shows amount of output values.

Scale Item: Values in matrix must be between 1 and -1 or between 1 and 0. If in Raw Data File, if values are not in this format, “scale item” option must be selected.

Minimum Output Value: It will be the minimum value of any row after scaling.

Maximum Output Value: It will be the maximum value of any row after scaling.

Ratio of Training: Neural network uses some data to train itself, and it uses some other for testing. This section fixes the ratio of train/test rows in data file.

Shuffle: It mixes rows each other.

Training: Program providing two options called “Normal” and “Cascade”. Normal training has an ANN with constant geometry. Cascaded training changes network geometry while training.

Stop Function: Criteria of ending training. Generally MSE (Mean Squared Error) is selected.

Max Epoch: Specifies number of the training steps.

FineTuning: Advanced options can be set in here.

Saving ANN: Program providing options while saving. We can save “Latest ANN” or ANN in “Minimum MSE Step”.

Training Methods: FannTool has 4 training methods.

FANN_TRAIN_INCREMENTAL is standard back propagation algorithm. It updates weights after any training data.

FANN_TRAIN_BATCH is also standard back propagation algorithm. It updates weights after looking all training data and calculating MSE.

FANN_TRAIN_RPROP improved by Riedmiller and Braun in 1993.

FANN_TRAIN_QUICKPROP improved by Fahlman in 1988.

4.4 TRAINING & TESTING

24 sections of CBC are inputs and Anemia result is output at first. This type will be called as Just-Anemia Network. Train & Test ratio is %50-%50. FannTool found FANN_TRAIN_INCREMENTAL as optimal training method. And also found FANN_ELLIOT as optimal hidden activation function and FANN_SIGMOID as optimal output activation function. After normal training with 3 layers (hidden layer has 11 weights) network, an ANN is ready for testing.

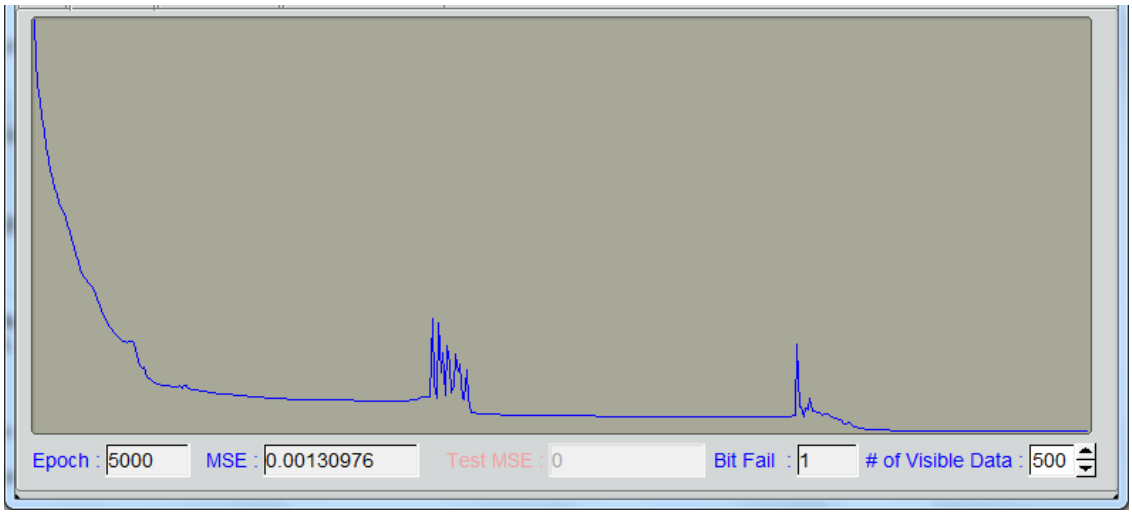

Figure 4.13 Just Anemia ANN training graph with MSE (0.0013).

Trained ANN tested with 803 data. 752 of them were Anemia and 51 of them were healthy. ANN resulted in 790 true, 13 false. All of false data are healthy samples. This means %98.4 total success, %100 Anemia sample is diagnosed successfully and %79.6 healthy sample is diagnosed successfully.Detailed results can be seen in Table 4.4 and training graph can be seen in Figure 4.13.

Table 4.4 Just Anemia ANN success table with MSE (0.0013).

Class	Correct	FALSE	Success %
Healthy	38	13	79.60%
Anemia	752	0	100%
Total	790	13	98.40%

Second design includes subtype of Anemia such as microcytic, normocytic or macrocytic (and also healthy). So, input is same but output has 4 different situations. This type will be called as Subtype Network. FannTool recommends FANN_TRAIN_INCREMENTAL for training method, FANN_GAUSSIAN method for hidden activation function, and FANN_SIGMOID_SYMMETRIC method for output activation function. At the and a 3 layers neural network (11 weights hidden layer) is designed with this functions and %50-%50 ratio for training & testing adopted.

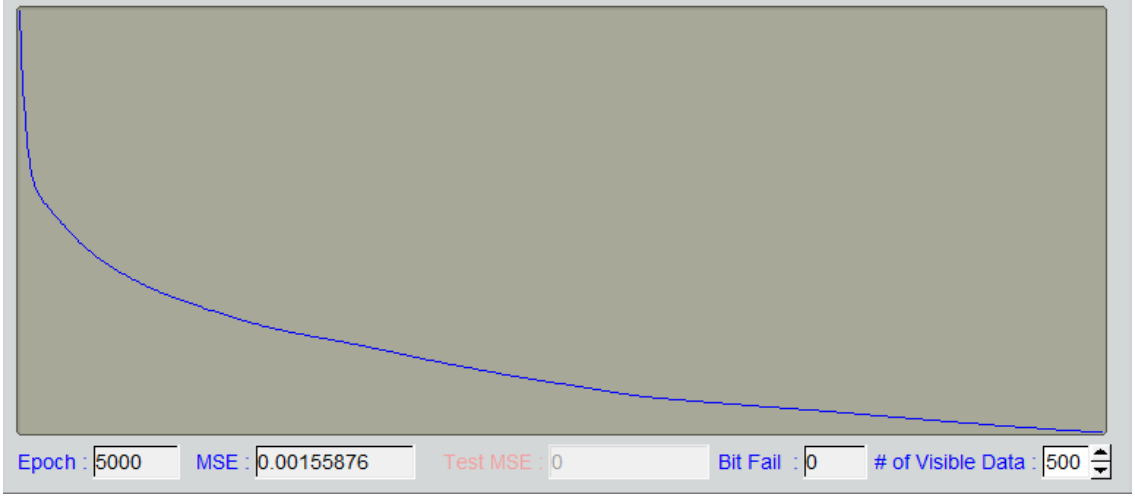

Figure 4.14 Subtype ANN training graph with MSE (0.0016).

At the end, ANN tests with 803 test samples. 58 of them healthy (57 true 1 wrong diagnosed) and others have some conditions. 48 of them have microcytic Anemia (46 true 2 wrong diagnosed), 101 of them (88 true 13 wrong diagnosed) have macrocytic Anemia and 596 of them (570 true 26 wrong diagnosed) have normocytic Anemia. Detailed results can be seen in Table 4.5 and training graph can be seen in Figure 4.14.

Table 4.5 Subtype ANN success table with MSE (0.0016).

Class	Correct	FALSE	Success %

Healthy	57	1	98.30%
Microcytic	46	2	95.90%
Normocytic	570	26	95.60%
Macrocytic	88	13	87.10%
Total	761	42	94.80%

Another experiment is executed with different network design. Inputs are same but outputs are different. There is no output representing different cases. There are 4 outputs represent microcytic, normocytic and macrocytic Anemia with healthiness. This type network will be called as 4-Section Network. A 3-layer 11-neurons network trained with INCREMENTAL training graph and results are below.

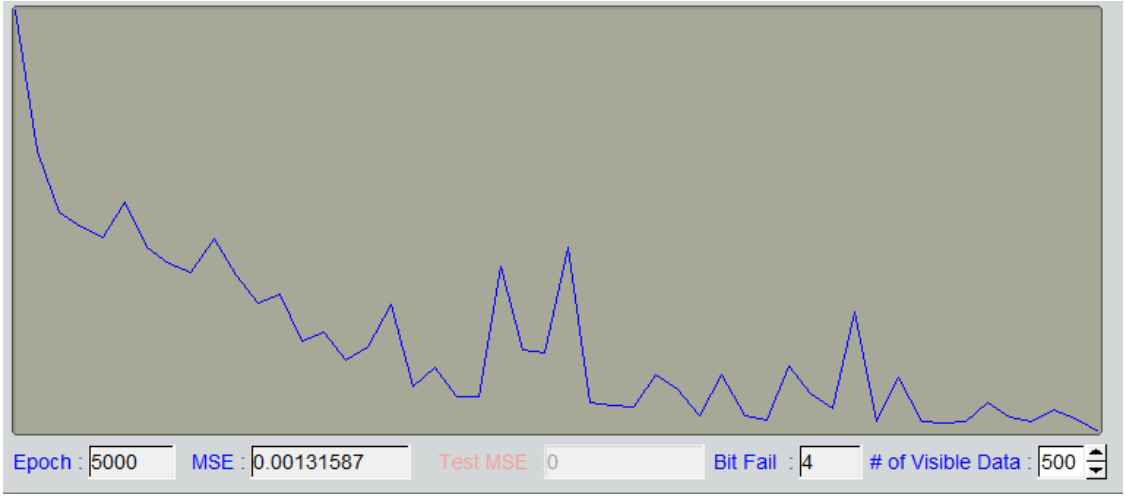

Figure 4.15 4-Section Network Training graph with MSE (0.0013)

There are 803 test samples, 742 Anemia and 61 healthy. 6 of healthy and 11 of Anemia was misclassified. And 40 samples in Anemia were diagnosed incorrectly in their Anemia subtype. Detailed results can be seen in Table 4.6 and training graph can be seen in Figure 4.15.

Table 4.6 Subtype ANN success table 2.

Class	Correct	FALSE	Success %
Healthy	55	6	90.20%
Anemia	731	11	98.50%
Subtype	702	40	94.60%
Total	786	17	97.90%
Total	757	46	94.30%

4.5 COMPARING DIFFERENT METHODS

In this research, effects of different methods on network is another important topic. Basically; changing train & test ratio and observing results, changing training method, activation function and observing results, switching from normal training to cascade training and observing results are studied in this topic.

4.5.1 Train Data Ratio and Number of Epochs

In FannTool, train & test ratio can be changed from 50-50 to 90-10. Our one-output 3-layer networks trained with using FANN_TRAIN_INCREMENTAL algorithm and after 5000 and 10000 epochs have these MSEs:

Table 4.7 Train and test ratio comparison table.

Ratio	5000 epochs	10000 epochs
50 % train – 50 % test:	0.00162	0.00095
60 % train – 40 % test:	0.00181	0.00119
70 % train – 30 % test:	0.00186	0.00118
80 % train – 20 % test:	0.00188	0.00146
90 % train – 10 % test:	0.00197	0.00138

As can be seen, if we increase train ratio of samples, MSE of trained network do not decrease. So, as it can be seen in Table 4.7, 50%-50% distribution of data looks suitable. Furthermore, we saw that number of epochs is significantly effective on MSE. 10000-epochs network provides 96 percent accuracy to classify test samples. It misclassified only 32 in 803 samples. If we continue to increase MSE, these MSEs will be seen.

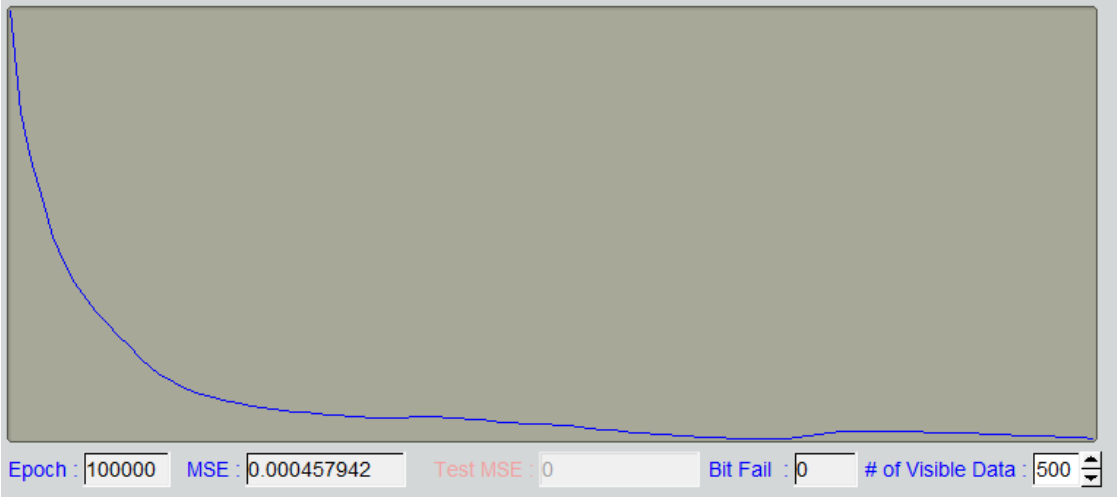

Figure 4.16 MSE changes in 100000 epochs.

Table 4.8 MSE success rates.

Epochs	MSE	Success %
10000	0.00095	96.00%
50000	0.00047	94.00%
100000	0.00045	93.50%

So, it can be seen in Table 4.8 that lack of MSE not always brings high accuracy and after 50000 epochs, MSE does not change significantly. It can be also seen in Figure 4.16 that 100000 epoch is too much for required training.

4.5.2 Training Method

FannTool found FANN_TRAIN_INCREMENTAL as optimal training method in just Anemia diagnose experiment. In Anemia subtype diagnosing test, FannTool recommended INCREMENTAL algorithm again. In this section, we search how MSE and success change, if network is trained by other algorithms. Epochs are 5000; ratio of training data is %50; activation function is SIGMOID.

Table 4.9 Success rates of different training algorithms.

Method	MSE	Success %
INCREMENTAL	0.00153	95.00%
BATCH	0.0086	85.30%
RPROP	0.00153	86.40%
QUICKPROP	0.00942	88.70%

In conclusion, INCREMENTAL algorithm, which FannTool recommends provides minimum MSE and best results. Detailed results are in Table 4.9.

4.5.3 Activation Function

FannTool has a suggestion tool for Hidden Activation Function and Output Activation Fuction combination. For finding optimum pair, it compares all combinations of activation function pairs. These are MSEs of these combinations with 2000 epochs and INCREMENTAL algorithm. This process took about 15 minutes on 2 GHz Intel Core 2 Duo processors.

Table 4.10 MSEs of activation functions.

Hidden A.F.	Best Output A.F. Pair	MSE
LINEAR	COS SYM.	0.0062
SIGM.	SIGM. SYM. STEP.	0.002
SIGM. STEP.	SIGM. SYM. STEP.	0.002
SIGM. SYM.	ELLIOT SYM.	0.0023
SIGM. SYM. STEP.	ELLIOT SYM.	0.0024
GAUSSIAN	SIGM. SYM.	0.0019
GAUSSIAN SYM.	ELLIOT SYM.	0.0022
ELLIOT	ELLIOT SYM.	0.0023
ELLIOT SYM.	ELLIOT SYM.	0.002
LINEAR PIE.	COS SYM.	0.0054
LINEAR PIE. SYM.	COS SYM.	0.0062
COS SYM.	ELLIOT SYM.	0.0046
SIN SYM.	COS SYM.	0.0044

In conclusion best pair is GAUSSIAN for hidden activation function and SIGMOID_SYMMETRIC for output activation function. Its MSE is 0.0019. Detailed results are in Table 4.10.

4.5.4 Number of Layers

In normal training, number of layers can be set. 3-layer network means it has one hidden layer. 4-layer means two hidden layers and 5-layer means three-hidden layers.

Another experiment executed with 4-layer network. It has 2 hidden layers with 17 neurons in hidden layer 1 and 10 neurons in hidden layer 2. It was trained with FANN_TRAIN_INCREMENTAL algorithm. Its training graph is below.

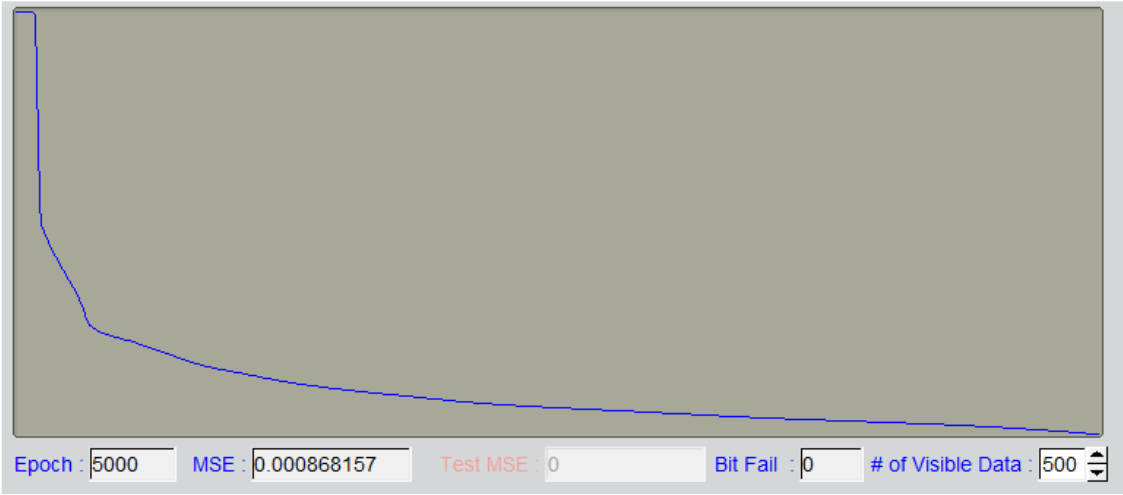

Figure 4.17 4-layer ANN training graph with MSE (0.0009)

This network's results are very impressive. In total, 769 of 803 data are successfully classified. Detailed results and training graph are in Figure 4.17 and Table 4.11.

Table 4.11 Success rates according to Anemia types.

Class	Correct	FALSE	Success %
Healthy	58	0	100%
Microcytic	48	0	100%
Normocytic	570	26	95.60%
Macrocytic	93	8	92.10%
Total	769	34	95.80%

Another experiment executed with 5-layer network with 19 neurons in hidden layer 1, 14 neurons in hidden layer 2, 9 neurons in hidden layer 3. It was trained with FANN_TRAIN_ALGORITHM as FannTool's recommendation. Training graph and results are below in Figure 4.18 and in Table 4.12.

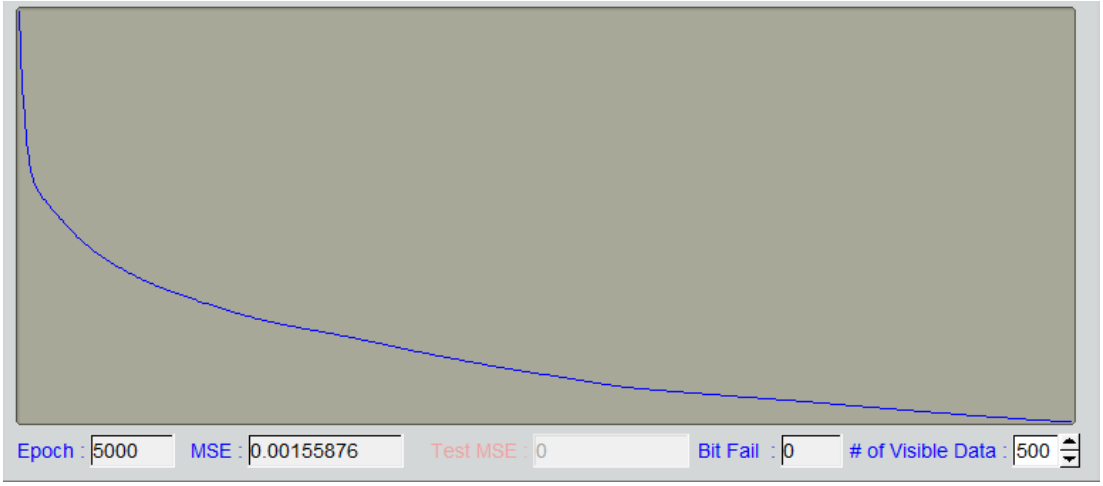

Figure 4.18 5-layer network

Table 4.12 Success rates of 5-layer network.

Class	Correct	FALSE	Success %
Healthy	58	0	100%
Microcytic	48	0	100%
Normocytic	596	0	100%
Macrocytic	0	101	0%
Total	702	101	87.40%

4.5.5 Normal vs. Cascaded Training

In normal training, node and weight number of network has to be set at first. Neural network improves its weights in this geometry. But in cascade training, neural network improves itself by updating both weights and network geometry. In this section, normal and cascade trained neural networks were compared.

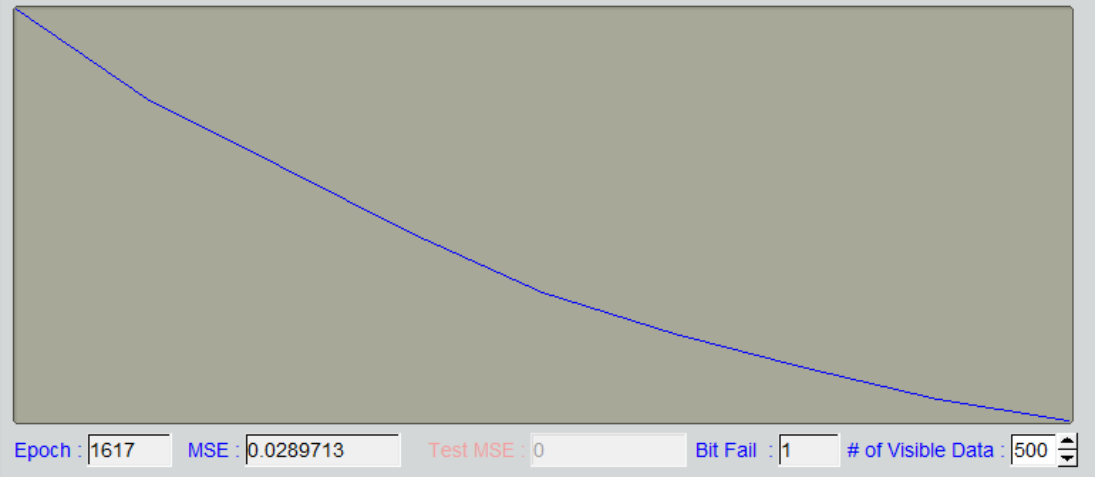

Figure 4.19 ANN (cascaded) training graph with MSE (0.0289).

MSE of network (~0.03) was too high to give acceptable results. As expected, network gave just 106 correct answers on 803 samples. But it can be seen in the training graph, MSE are still decreasing when training stopped. It means if we do not stop to train network, MSE will probably continue decreasing. If we look graph in Figure 4.19, we can see training was not stopped, so it explains failure on Table 4.13.

Table 4.13 ANN (cascaded) success rate table with MSE (0.0289).

Class	Correct	FALSE	Success %
Healthy	58	0	100%
Microcytic	48	0	100%
Normocytic	0	596	0%
Macrocytic	0	101	0%
Total	106	697	13.20%

This experiment was repeated at a satisfied number of epochs in training. Network was trained with FANN_TRAIN_RPROP algorithm and activation functions were FANN_SIGMOID_SYMMETRIC.

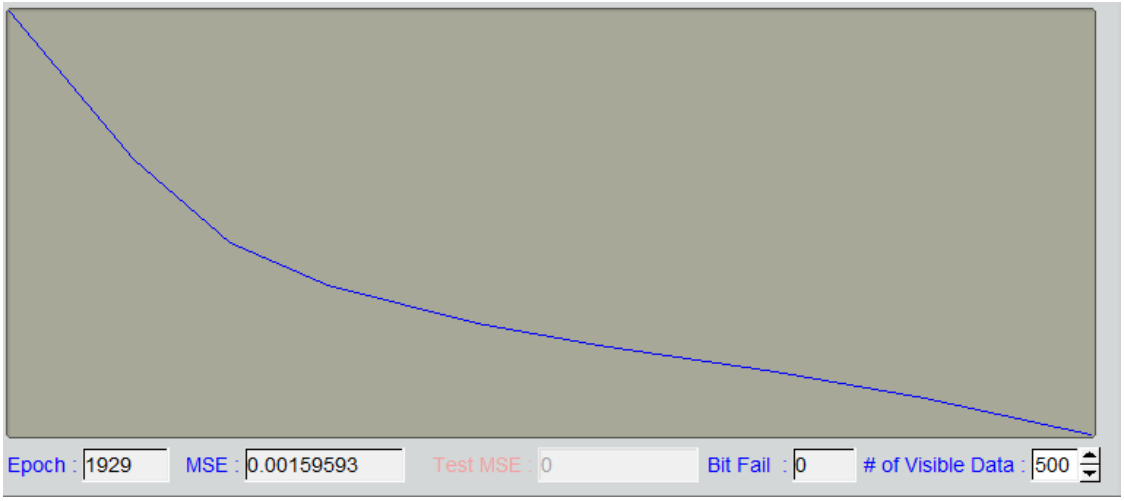

Figure 4.20 ANN (cascaded) training graph with MSE (0.0015).

These results are more satisfactory than previous network's results. Only 47 wrong classifications occurred in total 803 samples. ANN training graph can be seen in Figure 4.20 and success rate can be seen in Table 4.14. This time success rate is satisfactory with longer training-time.

Table 4.14 ANN (cascaded) success rate table with MSE (0.0015).

Class	Correct	FALSE	Success %
Healthy	55	3	94.80%
Microcytic	47	1	97.90%
Normocytic	561	35	94.10%
Macrocytic	93	8	92.10%
Total	756	47	94.10%

CHAPTER 5

CONCLUSION

5.1 EVALUATION OF RESULTS

In this thesis, medical data analyzed with comparing different ANN training algorithms like backpropagation, hopfield, perceptron, probabilistic etc. and comparing different training methods like linear, sigmoid, gaussian, elliot etc. Main purpose of this thesis is to make Anemia diagnosis from CBC results.

Required CBC data for training and testing ANNs is gathered from Istanbul University Istanbul Faculty of Medicine. 1606 samples are used in research 1278 of which are male and 328 of which are female. Samples' age distribution is from 18 to 49.

In the first step, since CBC results taken from İstanbul University has more than 20 elements, feature reduction becomes important. Then, 11 of elements found significant for Anemia diagnosis. These are: Gender, HCT, HGB, LYMPH#, MCH, MCHC, MCV, MPV, PCT, PLT, RBC and WBC. In addition, different training algorithms compared with MATLAB tool. Three training algorithms' success rates are over 90%. These are: Feed-Forward Distributed Time Delay (723 of 803), Layer Recurrent (727 of 803) and Cascade-Forward Back-Propagation (754 of 803).

Moreover, some different training methods like incremental, batch etc. and activation functions like linear, sigmoid, gaussian, elliot etc. are compared with Fast Artificial Neural Network (FANN) Tool. FannTool uses backpropagation algorithm and has a training method suggestion module and it suggest FANN_TRAIN_INCREMENTAL method. After comparing different training methods, we observed that the algorithm which is suggested by FANN Tool is the most efficient way because of providing minimum MSE (0.00153).

With changing train/test ratio and number of epochs, MSE changed significantly. Optimum train ratio found as 50% of total samples. (0.00162 MSE in 5000 epochs, 0.00095 MSE in 10000 epochs). MSE shows continued decline, when epochs are increased. (0.00047 in 50000 epochs, 0.00045 in 100000 epochs).

FannTool also has hidden activation function and output activation function suggestion module. For our samples, recommended functions by FannTool are GAUSSIAN for hidden layers and SIGMOID_SYMMETRIC for output layer. Combination has minimum MSE (0.0019) on 2000 epochs.

Also, ANNs with different layers are compared with FannTool. 4-layer ANN gave the most efficient result. It was 769 correct results on total 803 samples (95.8%).

In conclusion, diagnosing Anemia and subtypes of it from complete blood count analysis which is main purpose of this thesis is succeeded. With Backpropagation algorithm, Incremental training method, Gaussian-Sigmoid Symmetric activation functions pair with 4-layer neural network architecture gave us 769 correct results in 803 samples which means 95.8% correctness.

5.2 DISCUSSION

In this thesis, Anemia diagnosis with different types that are decided by mean corpuscular volume data coming from complete blood count was succeeded with ANN. Throughout this study, some difficulties are faced. One of them is, for Anemia subtype diagnosis; there is more than one blood test. For example, it is necessary to know iron

amount in blood for diagnosis iron deficiency Anemia. Iron level does not exist in CBC, so it cannot be diagnosed exactly with just analyzing this.

Following researches may be on complete Anemia diagnosis with analyzing more than one test results simultaneously.

REFERENCES

Allahverdi N., Tunalı A., Işık H., Kahramanlı, H., A Takagi-Sugeno type neuro-fuzzy network for determining child Anemia, 2011

Anderson, D. , McNeill, G., Artificial Neural Networks Technology, A DACS State-of-the-Art Report, 1992.

Collins, M., M.D., What is an Eosinophil?, 2004 , <http://www.apfed.org/downloads/What_is_an_Eosinophil.pdf>

Complete Blood Count, Lab Test Online, American Association for Clinical Chemistry (AACC), 2005, <http://www.aacc.org/Pages/default.aspx>

Dallman PR., Blood and blood-forming tissues. In: Pediatrics. Rudolph A, ed. 16th ed. E. Norwalk, CT, Appleton-Century-Crofts, 1977

Dematos G., Boyd M.S., Kermanshahi B., Kohzadi N., Feedforward versus recurrent neural networks for forecasting monthly japanese yen exchange rates, 1996

Dorlands Medical Dictionary: lymphocyte, 2007. Retrieved 2009-01-27.,<http://www.dorlands.com/wsearch.jsp>

Eisenstaedt R, Penninx BW, Woodman RC. Anemia in the elderly: current understanding and emerging concepts, 2006.

Guerrero, J. D. M., Olivas, E. S., Camps Valls, G., Serrano Lopez, A. J., Perez Ruixo, J. J., & Jimenez Torres, N. V., Use of neural networks for dosage individualization of erythropoietin in patients with secondary Anemia to chronic renal failure.Computers in Biology and Medicine, in press, 2003.

IBM and EPFL Join Forces to Uncover the Secrets of Cognitive Intelligence, 2005. Retrieved 2009-05-02.<https://researcher.ibm.com/researcher/switch_views.php>

International Classification of Diseases (ICD), World Health Organization (WHO), <http://www.who.int/classifications/icd/icd10updates/en/index.html>

Izhikevich E. M., Polychronization: computation with spikes, 2006.

Maton, Anthea; Jean Hopkins, Charles William McLaughlin, Susan Johnson, MaryannaQuon Warner, David LaHart, Jill D. Wright, Human Biology and Health., 1993.

MedlinePlus Medical Encyclopedia: RBC indices". Retrieved 2009-03-0, <http://www.nlm.nih.gov/medlineplus>

Mehrotra, K. , Mohan, C.K. , Ranka, S. Elements of Artificial Neural Networks, 1996.

National Health Nutrition Examination Survey (NHANES 2005-2006), <http://www.cdc.gov/nchs/nhanes.htm>

Predicting the effectiveness of hydroxyurea in individual sickle cell Anemia patients. Artificial Intelligence in Medicine, 2000

Purves, William K.; David Sadava, Gordon H. Orians, H. Craig Heller, Life: The Science of Biology(7th ed.), 2004.

Sprogar, M., Lenic, M., &Alayon, S., Evolution in medical decision making, Journal of Medical Systems, 2002

Swirski FK, Nahrendorf M, Etzrodt M, Wildgruber M, Cortez-Retamozo V, Panizzi P, Figueiredo J-L, Kohler RH, Chudnovskiy A, Waterman P, Aikawa E, Mempel TR, Libby P, Weissleder R, Pittet MJ., Identification of Splenic Reservoir Monocytes and Their Deployment to Inflammatory Sites, 2009.

Valafar F, Ersoy OK. PNS modules for the synthesis of parallel self-organizing hierarchical neural networks.Internl J Circuits, Syst Signal Processing 1996

Valafar, H., Valafar, F., Darvill, A., Albersheim, P., Kutlar, A., Woods, K.F., Hardin, J. Van Beekvelt MC, Colier WN, Wevers RA, Van Engelen BG. "Performance of near-

infrared spectroscopy in measuring local O2 consumption and blood flow in skeletal muscle", 2001

Witko-Sarsat, V; Rieu P, Descamps-Latscha B, Lesavre P, Halbwachs-Mecarelli L., Neutrophils: molecules, functions and pathophysiological aspects, 2000.

Wiwanitkit V. Plateletcrit, mean platelet volume, platelet distribution width: its expected values and correlation with parallel red blood cell parameters. ClinAppl Thrombosis Hemostasis 2004

Zini, G. and d'Onofrio, G., Neural Network in Hematopoietic Malignancies, 2003.