JANUARY 2018 M.Sc. in Electronics and Computer Engineering ARAS AHMED ALI

UNIVERSITY OF GAZİANTEP

GRADUATE SCHOOL OF

NATURAL & APPLIED SCIENCES

KNOWLEDGE DISCOVERY IN HEALTH DOMAIN USING DEEP NEURAL NETWORK ALGORITHMS

M. Sc. THESIS

IN

ELECTRONICS AND COMPUTER ENGINEERING

BY

ARAS AHMED ALI

JANUARY 2018

Knowledge Discovery in Health Domain Using Deep Neural Network Algorithms

M.Sc. Thesis

In

Electronics and Computer Engineering

University of Gaziantep

Supervisor

Prof. Dr. Ergun ERÇELEBİ

By

Aras Ahmed ALI

January 2018

REPUBLIC OF TURKEY
UNIVERSITY OF GAZIANTEP
GRADUATE SCHOOL OF NATURAL & APPLIED SCIENCES
ELECTRICAL AND ELECTRONICS ENGINEERING

Name of the thesis: Knowledge Discovery In Health Domain Using Deep Neural Network Algorithms
Name of the student: Aras Ahmed Ali
Exam Date: 17/01/2018
Approval of Graduate School of Natural and Applied Sciences

Prof. Dr. Ahmet Necmeddin YAZICI
Director

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

Prof. Dr. Ergun ERÇELEBİ
Head of Department

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

Prof. Dr. Ergun ERÇELEBİ
Supervisor

Examining Committee Members	Signature
Prof. Dr. Ergun ERÇELEBİ	.....
Prof. Dr. Ilyas EKER	.....
Assoc. Prof. Dr. AHMET METE VURAL	.....

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and proper conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Aras Ahmed ALI

ABSTRACT

KNOWLEDGE DISCOVERY IN HEALTH DOMAIN USING DEEP NEURAL NETWORK ALGORITHMS

ALI, Aras Ahmed Ali

M.Sc. in Electronics and Computer Engineering

Supervisor: Prof. Dr. Ergun ERÇELEBİ

January 2018, 37 Pages

With the nowadays technology and the plenty of information in the health care system, there is a need to extract a useful knowledge that can be used to diagnose and identify patterns in the patient records. The process to extract such knowledge is called data mining. The steps of pattern extraction and discovery involve a complex process, which normally uses a large amount of datasets. There are several application and systems implemented to diagnosis patient records in health care and clinical data.One of them is called the knowledge discovery in database (KDD) which mostly depends on developing a method that can process the data in a good manner. Data mining steps are considered one of crucial steps in KDD process in order to extract a useful pattern from the dataset. In order to achieve better healthcare services, data mining requires a proper design and implementation of data mining algorithm to identify a unique pattern from the data. In this research we suggest using Patient Information for the Hewa Hospital in Sulamani, which is responsible for the cancer and blood decease as a case study. The main aim of this study is to investigate the deep neural network (DNN) and Artificial neural network (ANN) as classification algorithms in order to help us for better decisions. The obtained results demonstrate that, the DNN achieves better performance in compare to Artificial Neural Network (ANN) algorithm. The scored results reaches up to 87.84 when using 70:30 training and testing dataset

Key words: Deep neural network (DNN), artificial neural network (ANN), knowledge discovery.

ÖZET

DERİN SİNİR AĞI ALGORITMALARI KULLANARAK SAĞLIK ALANINDA BİLGİ KEŞFİ

ALI, Aras Ahmed Ali

M.Sc. in Electronics and Computer Engineering

Danışman: Prof. Dr. Ergun ERÇELEBİ

Ocak 2018, 37 Sayfa

Günümüz teknolojileri ve sağlık sistemindeki bol miktarda bilgi ile hasta kayıtlarındaki kalıpları teşhis etmek ve tanımlamak için kullanılabilecek yararlı bir bilgiyi çıkarmak gerekir. Bu bilgiyi elde etme işlemi veri madenciliği olarak adlandırılmaktadır. Desen çıkarma ve keşif adımları, genellikle çok miktarda veri kümesi kullanan karmaşık bir süreç içermektedir. Sağlık kayıtlarında ve klinik veride hasta kayıtlarını teşhis etmek için uygulanan birçok uygulama ve sistem vardır. Bunlardan birine, veriyi iyi bir şekilde işleyebilecek bir yöntem geliştirmeye dayalı olan, veritabanındaki bilgi keşfi (KDD) adı verilmektedir. Veri madenciliği adımları, veri kümesinden yararlı bir model çıkartmak için KDD sürecinde önemli adımlardan biri olarak değerlendirilmektedir. Daha iyi sağlık hizmetleri elde etmek için, veri madenciliği veriden benzersiz bir model belirlemek için veri madenciliği algoritmasının uygun bir tasarım ve uygulanmasını gerektirir. Bu araştırmada, kanser ve kan kaybından sorumlu olan Sulamani'deki Hewa Hastanesi için bir Hasta Bilgilendirme Yöntemi kullanılmasını öneriyoruz. Bu çalışmanın temel amacı daha iyi kararlar almamıza yardımcı olması için derin sinir ağı (DNN) ve Yapay sinir ağının (YSA) sınıflandırma algoritmaları olarak araştırılmasıdır. Elde edilen sonuçlar DNN'nin Yapay Sinir Ağı (ANN) algoritmasına kıyasla daha iyi bir performans elde ettiğini göstermektedir. 70:30 eğitim ve test veri seti kullanıldığında puanlanan sonuçlar 87.84'e ulaşmaktadır.

Anahtar kelimeler: Sağlık, Derin sinir ağı (DNN), Yapay sinir ağları (YSA), Bilgi Keşfi

To the “values” that were always “relevant”

And never “redundant” in my life

My parents, my sisters, my brothers

ACKNOWLEDGEMENTS

First, I am grateful to The Almighty God for establishing me to complete this work there are many people who encouraged me in the development stage of this thesis during the last years. First of all I want to express my Warmest thanks and gratitude and appreciation to my supervisor Prof. Dr. Ergun ERÇELEBİ for this Sincere supervision of this work and for this valuable suggestion

I owe a debt to Dr.Anas Younis Hamza, Dr.Ayser Abdulkhaleq Abdulrahman in university of Sulaimani PhD computer engineering, and Dr.Alaa K Jumaa head of database department in sulaimani polytechnic university, for providing necessary facilities, and for their Support and Special thanks to Dr.Mardin Mahsum Faraj who help to during my study.

I also want to thanks HEWA Hospital, gave me the opportunity to continue this work at his Hospital found a very positive and Supportive environment giving me much freedom for my research. I want to thank my parents for their affection and their help for Managing my life in busy times. I want to thank the rest of my family and My wife's parents. In addition, my sincere thanks to my wife, SAKAR for all her Love and support.

Finally, I would also like to thank my friends Eng. Khabat Star and Payam Ali Mohammed, who helped me a lot in finalizing this work within the limited period

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURE

LIST OF ABBREVIATIONS

DNN	Deep Neural Network
ANN	Artificial neural network
SVM	Support vector machine

CHAPTER 1

INTRODUCTION

1.1 Background

The rapid growth of the dataset sizes and technology requires knowledge discovery for health domain, which intends to have a huge impact on the health care system in the future. Recently, the world has witness a fast increase of population. As well as, the cost of the health care is increasing dramatically. One way to handle these changes is to improve the process of healthcare and medical care using the power of data mining and technology. Moreover, there is a high demand to improve the healthcare systems to support the patients in several aspects

In the health domain, one of the most significant key of using data mining is to extract hidden relationships within a huge amount of dataset. It is very hard to extract hidden information associated with one to another on only one record by an ordinary and known analysis method [1].

For instance; if we assume that there is a patient suffers from high blood pressure, the evaluating of one variable is not accurate because the blood pressure is affected by many other variables in the human body. Therefore, the analysis supposed to include several variables to achieve the most accurate results. The analysis of the huge data is a challenging task and requires a great knowledge to the dataset structure and implemented technique for knowledge discovery [2].

1.2 What is Data Mining?

Data mining refers to extracting or "mining" knowledge from large amounts of data. This terminology was first formally put forward by Usama Fayyad at the First International Conference on Knowledge Discovery and Data Mining, held in

Montreal in 1995 and still considered one of the main conferences on this topic. Many other terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. Many people treat data mining as a synonym for another popularly used term, "Knowledge Discovery from Data, or KDD [25, 26]. Data mining is a new field at the frontiers of statistics and information technologies (database management, artificial intelligence, machine learning, etc.)Of computer-assisted this aims at automatically mine large volumes of integrated data for new, hidden or unexpected information, or patterns in large data sets [27, 28]. It's not just about the use of a computer algorithm or a statistical technique; it is a process of business intelligence that can be used together with what is provided by information technology to support company decisions. It is to allow a corporation to improve its marketing, sales, and customer support operations through a better understanding of its customers. Data mining is a term that covers a broad range of techniques being used in a variety of industries Data mining is a relatively new field in both research and practice

1.3 Knowledge Discovery and Data Mining

Knowledge discovery and data mining has recently emerged as an important research direction for extracting useful information from vast repositories of data of various types. Before one attempts to extract useful knowledge from data, it is important to understand the overall approach. Simply knowing many algorithms used for data analysis is not sufficient for a successful data mining (DM) project Therefore, this Chapter focuses on describing and explaining the process that leads to finding new knowledge. The process defines a sequence of steps that should be followed to discover knowledge (e.g., patterns) in data. In fact useless to start the discovery of useful information from the data if they obtained result is not understandable by the user. This chapter discusses some of the basic concepts and issues involved in this process with special emphasis on different data mining tasks [29,30].

1.4 Knowledge Discovery Process Models

The efforts to establish a Knowledge Discovery in Databases (KDD) model were initiated in academia. In the mid-1990s, when the DM field was being shaped, researchers started defining multi-step procedures to guide users of DM tools in the complex knowledge discovery world. The main emphasis was to provide a sequence of activities that would help to execute a KDD, The process model developed in 1996 is the six-step model

In recent years, using data mining techniques have gain many researcher attentions in the information industry [3]. The main process of knowledge discovery consists of several processes such as data cleaning, data integration, data selection, knowledge presentation. The main role of using data mining to any data analysis is to discover a pattern from there datasets. In addition, data mining is considered as one of the most important area of research in order to extract meaningful information from huge data sets [4].

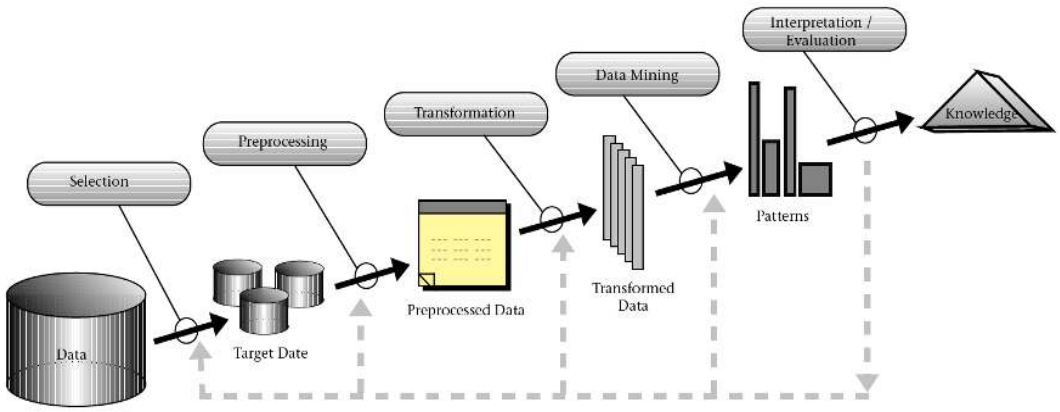


Figure 1.1 knowledge discovery main steps

Nowadays, the implementation of data mining methods and techniques is becoming popular in healthcare field due to the need of develop an effective methodology to predict the future outcome in health domain area [5]. Data mining techniques are implemented on several domains such as fraud detection for health, identify diseases, relevant medical treatments for the patient and find the cheapest cost for patients.

Moreover, it also assists the researchers' community in healthcare to create and build policies and drug recommendation systems for patients. The health care organization issues a data that is a very rich with information. This information contains with very complex information and factors and the process of extracting a pattern is one the most challenging task in health care domain [6].

Based on the above literature review we can find that the knowledge discovery in health domain requires a powerful technique to extract all the relevant information. Also, the analyses of health care systems are able to enhance the performance of management task and increase the efficiency. In terms of categorize the patients based on their diseases or illness type [5] [7].

In health domains, the data are characterized with a high dimensional and highly distributed with many values. Therefore, the power of data mining aims to discover hidden patterns to understand more relationship from the dataset [7].

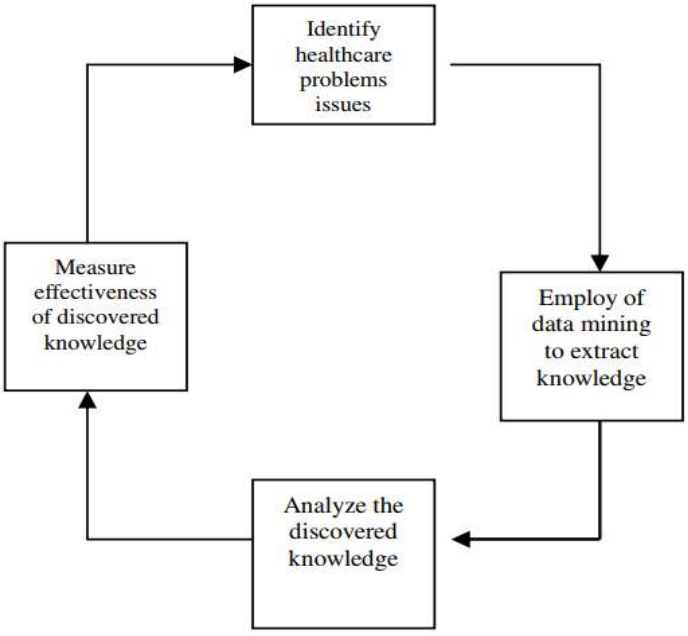


Figure 1.2 Health care data mining cycle.

Massive healthcare data needs to be converted into information and knowledge, which can help control, cost and maintains a high quality of patient care. Healthcare data includes patient-centric data and aggregate data.

1.4.1 Data Selection

As a first step the data needs to be carefully selected. Selection criteria include e.g. data availability, quality, type and format. The selection of high quality data semantically corresponding to the goal of the discovery process is essential for the following steps [31].

1.4.2 Data preprocessing

Is an important step in the knowledge discovery process: real world data are susceptible to noisy, missing and inconsistent data? Furthermore, the most of the information, if not regularly controlled and selected, can lead to over time consuming mining computational steps and to obtain answers liable to inaccuracy or mistakes. This step is particularly important when the considered domain is dynamic and liable to sudden changes. Pre-processing is known as the combination of steps needed to resolve each of the problems presented above and it takes place before the actual mining phase.

Data pre-processing is composed of several steps depicted in Figure (1.3) are: Data Cleaning, Data Integration, Data Transformation and Data Reduction.

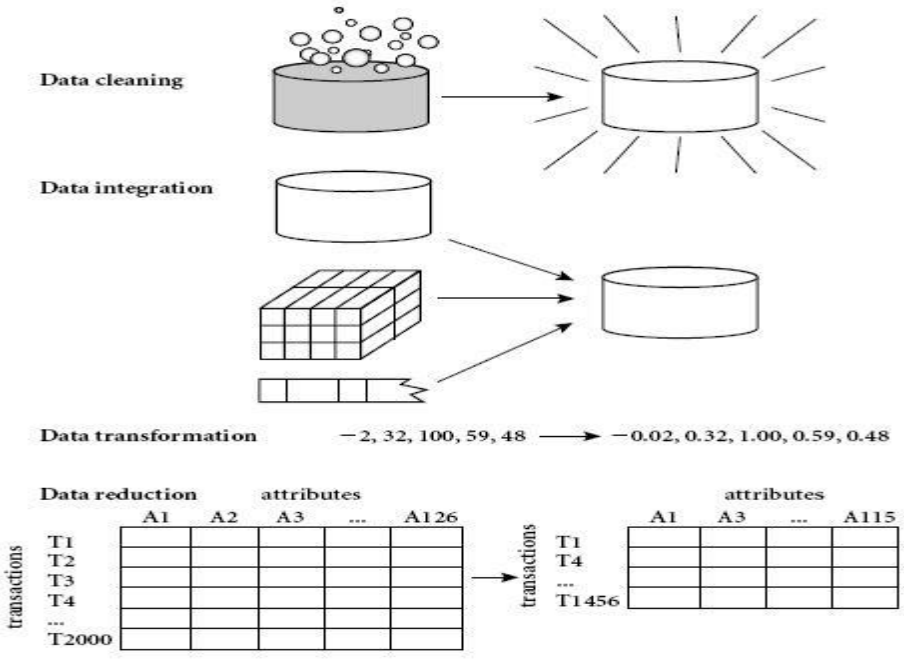


Figure 1.3 Data preprocessing

1.4.3 Data Transformation

Another important pre-processing task is data transformation. In data transformation, the data are transformed or consolidated into forms appropriate for mining. The researcher will review a few general types of transformations of data that are not problem-dependent and that may improve data-mining results. Selection of techniques and use in particular applications depend on types of data, amounts of data, and general characteristics of the data-mining task. The most common transformation techniques are Aggregation, Generalization, and Normalization [30, 33 and 34].

1.4.4 Data mining

The application of efficient algorithms to detect the desired patterns contained within the given data. Thus, the data mining step is responsible for finding patterns according to the predefined task. Since this step is the most important within the KDD process [32].

1.4.5 Evaluation Interpretation

At last, the user evaluates the extracted patterns after applying data mining techniques in a model with data sets, the result of the model will be interpreted. In order to properly interpret knowledge patterns, it is important to choose an appropriate visualization tool. If the user is satisfied with the quality of the patterns, the process is terminated.

However, in most cases the results might not be satisfying after only one iterations In those cases, the user might return to any of the previous steps to achieve more useful results. Besides the result of the model, some evaluation criteria should be taken into account. Statistically significant patterns can be found. The data interpretation stage is very critical; have to be careful as to how one interprets the patterns presented to us by data mining algorithms [35, 36, and 37].

1.4.6 Knowledge Presentation

Presentation of the information extracted in the data mining step in a format easily understood by the user is an important issue in knowledge discovery. Since this

module communicates between the users and the knowledge discovery step, it goes a long way in making the entire process more useful and effective. Important components of the knowledge presentation step are data visualization and knowledge representation techniques. Many visualization packages and tools are available, including pie charts, histograms, box plots, scatter plots, and distributions [29, 38].

1.5 Research Questions

As stated earlier. There are drawbacks and weakness of in terms of knowledge discovery for healthcare domain. Therefore, the main research question can be

- How can develop an Algorithm to that extract a pattern for healthcare data set?

- What is the impact of this algorithm to enhance the performance of healthcare domain?

1.6 Research Objective

This research focuses on the development of scalable and effective algorithms for healthcare data.

- To develop a deep neural network to extract a useful pattern for health care domain.

1.7 Research Scope

The main aim of this thesis is to develop a method that aims to classify the cancers with respect of their categories based on their medical record using deep neural networks (DNN). This research focuses on enhancing the performance of the previous method by implementing the deep neural network for healthcare dataset. In addition, the aim of this research is to extract a pattern from healthcare data.

1.8 Thesis Outlines

This thesis consists of five chapters; the first chapter intends to introduce background of this study, problem statements and research scope and the methodology of this research.

Chapter II introduces the previous related studies to have better understanding for the research problem and the techniques that have been implemented in this domain As well as a holistic overview of health care issues using data mining techniques and tools

Chapter III explains the methodology steps of this research in details. The methodology has many steps which are the development, improvement and evaluation phases. In addition, there is a details explanation for the dataset pre-processing steps and the tools that will be used in this research.

Chapter IV the chapter presents the experimental tests results along with the evaluation explanation that aims to evaluate the effectiveness of the proposed algorithm.

Chapter V provides the conclusion of the research. It provides the conclusion of this research and highlights the major contribution of this research with the some recommendation and suggestion for a future work and improvement.

CHAPTER 2

LITERATURE REVIEW

2.1 BACKGROUND

This chapter introduces several fundamental of classification algorithms in health domain. The review evidently explores the implementation of the algorithms for classification, this chapter concentrate on the state of the art methods and techniques in the literature to provide a general understanding of the tools and techniques in this study.

There are many different choices for constructing each classifier of the chain. So far, the same classifier for all classifiers in the chain among the possible choices, this work, the following classifiers:

2.2 Naive Bayes classifier

The Naive Bayes algorithm has been used in different classification healthcare systems in order to predict the performance of the dataset. The strength of the Naïve Bayes is that it has the ability of computing the patient type probability based on their diseases or sickness categorizes into a different classes [8].

The main significant of the Naïve Bayes algorithm is that known with an easy implementation for classification and achieves a very strong and high accuracy [9]. The data normally belongs to different classes, the naïve Bayes mainly deepened into classify the instances based on their classes.

Let's take an example as follows: given a patient P which indicates that there will be a group of features that are belonging to the patient $\begin{cases} d \\ i \end{cases} f = a, b, \dots\dots.(p)$ and P_c is denote the patient.

Where every patient will be categorize in a different set. The X denotes the conditional probability c for a patient P Where P_c means the number of the corresponding classes for each patient in the dataset [10, 11].

$$p(c_i \mid d_j) = \frac{p(c_i) . p(d_j \mid c_i)}{p(d_j)} \quad \text{……………………………………… ……………….. (2.1)}$$

Where d_j the number of testing patients and c_I is represents the category of each patient. The subsequent probability of every category c_I represented in the testing patient documents d_j, i.e. $P(c_I \mid d_j)$, is considered and the highest probability category is allocated to d_j. To calculate $P(c_I \mid d_j)$, $P(d_j \mid c_i)$ and $P(c_i)$ the evaluation rely on the training set. $P(d_j)$ is same for each category so we may remove it from the calculation.

2.3 K-nearest neighbor classifier

The K-nearest neighbour (KNN) is of one of the most common classifiers that rely on the category of the patient. Where every patient will be categorized based on a different class label. The learning process in the KNN classifier depends on the provided labels from the training dataset. The KNN stores all the patient records based on its similarity using a distance function. This distance function estimates a pattern from the dataset with respect the patient medical record and the other features that are in the system [12].

Given a medical record R, the objective function attempts to find the similarity K nearest neighbours for each patient using their medical history record. The measurement of the similarity score for each patient depends on the testing data sets and the number of class labels for the patients. The measurement formula that calculates the score in KNN as follow:

$$\text{score}(r, f_i) = \sum_{r_j = KNN(r)} sim(r, r_j)\, \delta(r_j, r_i) \quad \text{……………………………………….. (2.2)}$$

The KNN(r) represents the total number of nearest neighbors that each patient belonging R. Where r_j categorize as class of f_i, $\delta(r_j, f_i)$ which will be in two main values either (1 or 0). In the testing data, R will be match to the class that score the best result from the training set [13].

2.4 SVM classifier

Support Vector Machines (SVM) is one of the supervised algorithms that widely implemented for classification and regression analysis. The support vector machines are the most effective algorithms that categorize the training data into the belonging classes with respect to their classes to find the optimal solution in the search space [14]. SVM uses the probabilistic classification setting to model the solution, which mainly classifies the classes' problem in the hyper-plane of data into several points.

Let's assume that Y is a group of features which represented in points (y_1, z_1), (yn, z_n), the training point $y_i \in R^N$ is given a label $z_i \in \{-1, +1\}$, where i = 1,. . .,n. By using the training data, the measure process of finding hyper plane is computed using the following equation:

$$w \cdot y + b = 0 \text{ and } \begin{cases} w \cdot x_i + b \geq +1 \text{ if } z_i = +1 \\ w \cdot x_i + b \leq -1 \text{ if } z_i = -1 \end{cases} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.3)$$

In the above equation, *i* represents the number of feature points *N*, and the dot (.) represents the $w \cdot y = \sum_i w_i \cdot y_i$, that is used for each feature vectors *w* and *y*. In the learning process, the SVM attempts to increase the exploration in the search space to reach the optimal separating hyper-plane (OSH). The OSH has a maximal margin for both sides, which can be measured as the equation below:

$$\text{Minimize}(w) = \frac{1}{2}||w||^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.4)$$

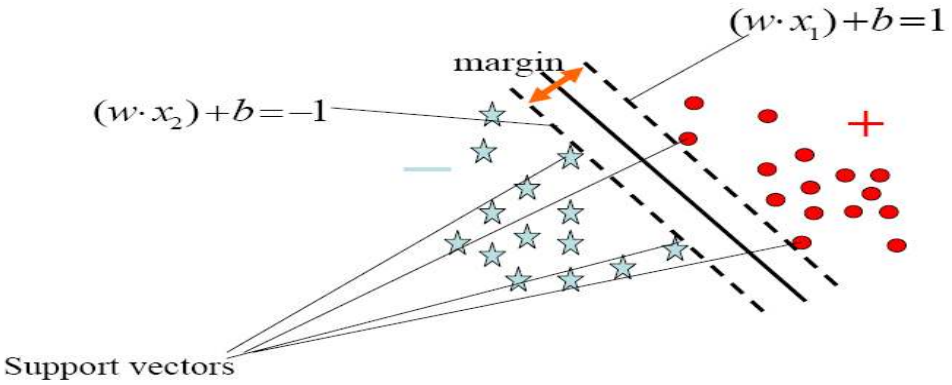


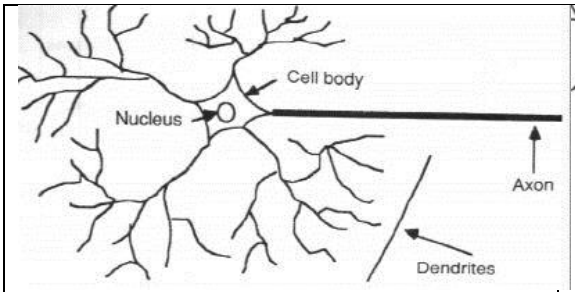
Figure 2.1 (SVM illustration).

In figure 2.1 Small circles, small stars and lines represent the examples of positive training, examples of negative training and decision surfaces, respectively. In this work, the support vector machine library (LibSVM) is adopted. LibSVM needs the training data to be in a specific format. Therefore, this work implements several functions that convert training data to this format. Figure 3.10 show example of the training data representation [15].

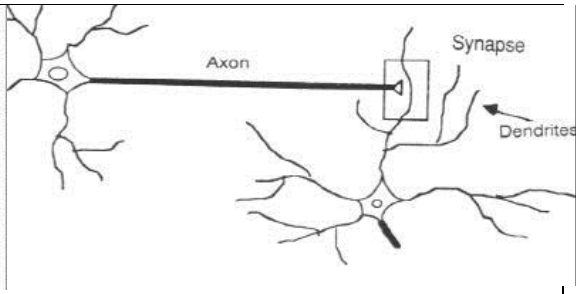
2.5 Artificial Neural Network (ANN)

An Artificial Neural Network (ANN) is one of the most computational model based on biological nervous systems. The ANN has been used in several real world application for example the prediction systems, banking and industry. The information that is provided to ANN network has a huge impact of the structure because the neural network learns and changes based on the amount of the input information. One of the ANN advantages is that the neural network can learn from the observed datasets using the random function.

The ANN main structure has three layers which are the (input layer, hidden layer and output layer) to process the data and extract a useful pattern. During the learning process, the process of the system needs adjustments for synaptic connections which are links the neurons [16].



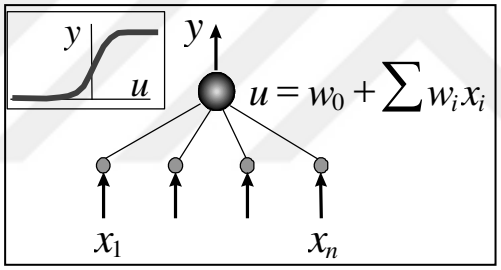
Components of a neuron



The synapse

Figure 2.2 neuron structure

In addition, the ANN model relies on the neuron weights, whereas the weighting schema affects the input data in the neurons, which leads to improve and increase the learning process in the ANN model [21].



$$y = f\left(\sum_{j=1}^{d} w_j x_j + w_0\right) \equiv f\left(\sum_{j=0}^{d} w_j x_j\right) \tag{2.5}$$

The calculation of the neurons are able to change based on the input weights which makes threshold very efficient and effective.

A. Feedback networks

Feedbacks networks are presented in the figure below, which have signals that can moves in two directions like a loop between the networks itself. The feedback networks have dynamic characteristics it keeps moving to till in reach to the stratification point in the network.

.

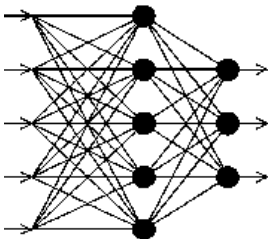


Figure 2.3 feedback neural net work

The state of the neural network remains stable until there is some changes in the inputs which is require to changes of the structure again

2.6 Deep Neural Networks

A deep neural network is a neural network with a certain level of complexity, a neural network with more than two layers. Deep neural networks use sophisticated mathematical modeling to process data in complex ways

Deep neural networks (DNN) have the exact same structure as other type of neural networks, except that the number of hidden layers is greater so that they can be considered deep

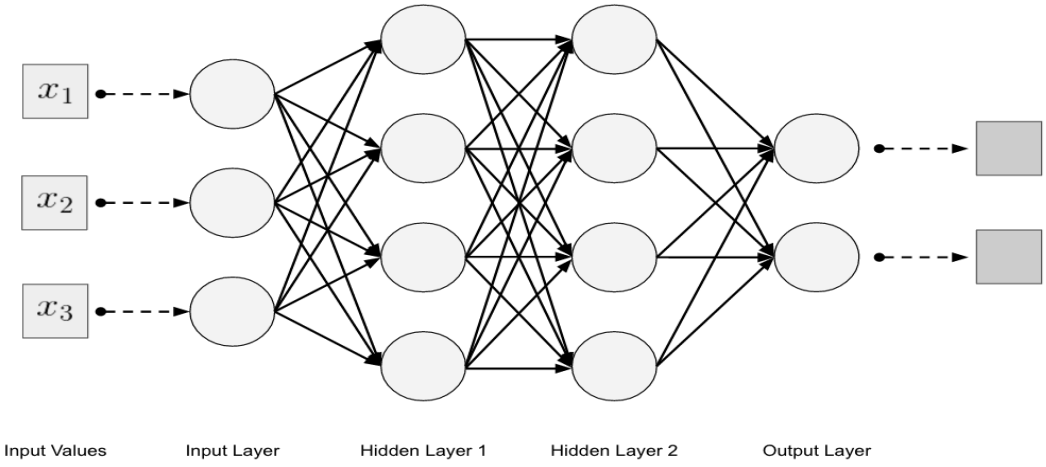


Figure 2.4 Deep Neural Network

CHAPTER 3

MATERIAL AND METHODS

3.1 INTRODUCTION

This chapter describes the research methodology of this research which is used to pre-process the dataset and develop the deep neural network for healthcare dataset. Moreover, this chapter provides a general step such as the research method phases, experimental design explanations and the techniques that are used in this experiment for the evaluation process.

The main aim of the literature review chapter is to understand the main concepts, definition and methodology for the previous studies to find the research gap. The theoretical part of this study in the previous chapter summarized the main definitions and concepts to identify the research gab for this study. Followed by, the main concept of the classification methods which mainly reviews possible improvements for healthcare data.

3.2 Research Method

This section provides more details of the main steps in order to classify the health medical data which including data pre-processing, classification and evaluation.

Firstly, the pre-processing techniques are implemented to clean and prepare the data form all the noisy and irrelevant data. This step is considered an important step for the implementation to ensure that the data are suitable and prepared as an input for the classification. Secondly, classification methods used for training and testing

To the healthcare dataset finally, the evaluation phase uses different data sizes to check performance using different data sizes.

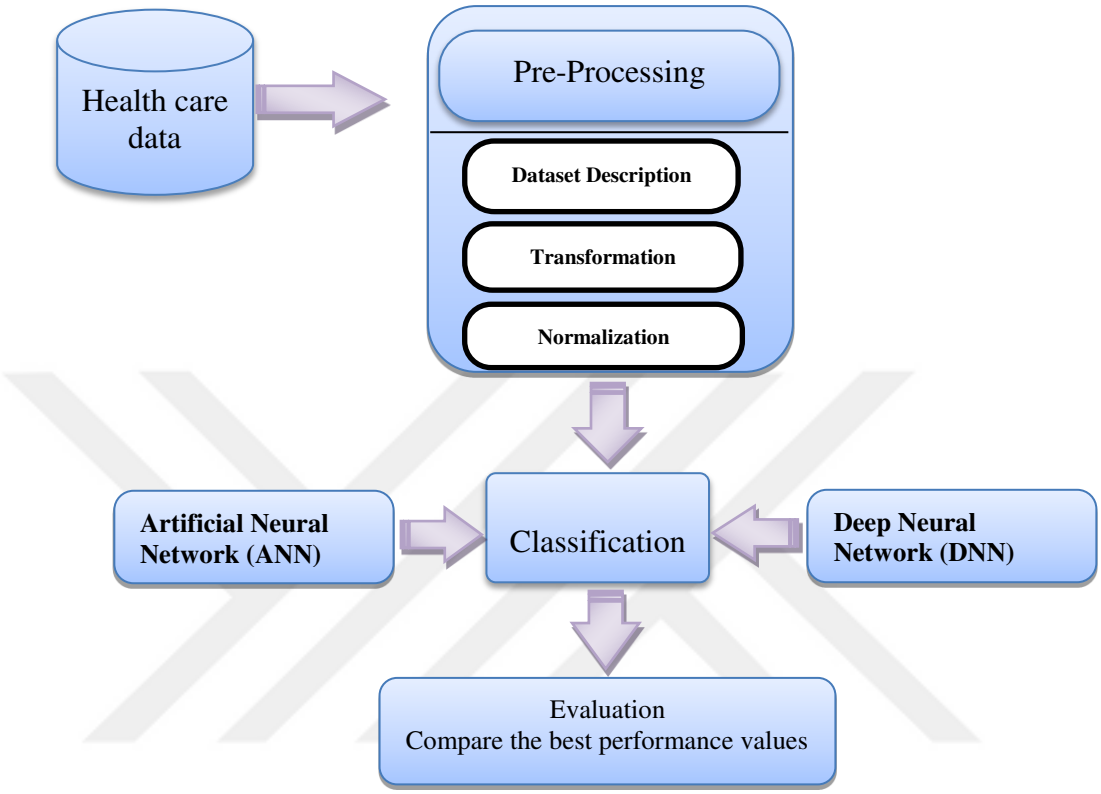


Figure 3.1 Health care System Architecture

3.2.1 **Data Pre-Processing**:

In this section, the data preprocessing steps is explained in details. This dataset contains all the information to determine an accurate decision that helps the hospital staff to reach better patient-related decisions. Once data preprocessing techniques aim to clean the dataset and enhance the accuracy and proficiency of the further mining process.

In fact, the data pre-processing considered one of the most significant steps, the reason behind that is that allows and assist in extracting in knowledge discovery.

a. Dataset Description

This big data set includes a lash amount of information that collected from an Iraq hospital for the patients. Table 3.1 demonstrates a short description of the dataset.

Table 3.1 data description

NAME OF DATA SET	DETAILS
Description	This data set contains patient's information.
Dataset characteristics	Categorical
Attribute characteristics	Integer or Categorical
Associated tasks	Classification, association rules
Number of instances	26199
Attributes number	9
Missing values	0
Area	Healthcare

Having an overall portrait of the data is very necessary to a successful data preprocessing. Descriptive data summarization techniques are useful for when applied to recognize your data type and distinguish noise or outliers data. Table 3.2 below shows the total number of the different categories found in the dataset.

Table 3.2 Attribute details

No.	Titles	Number	Details
1	Diagnosis Sub-Type	148	There are 149 diagnosis subtypes
2	Diagnosis Type	173	The total number of diagnosis is 174
3	Department	3	There are three departments which are Benign Hematology, Malignant Oncology, Malignant Hematology
4	Government	24	A total number of 25 governmental cities.
5	Patient jobs	40	The patient jobs are 41 founded in this dataset.
6	Blood Group	9	The blood group of the patients such as A+ O+, O-, B+, None, B- , AB+, A- AB-
7	Gender	3	The gender of the patients male , female or none
8	Department name	5	The department name are categorized into 5 main groups which are :- Pediatric , Hematology, Oncology, HLA Donor and BMT Patient)
9	Type	2	To describe the patient status Inpatient or outpatient

Once you have a lash amount of data and your data set is huge, data reduction technique should be applied to reduce the data set. Therefore, a normalization technique is used to decrease the Sparse of the dataset. At first, we count the distinct category for each attribute. The main aim of classifying the data is to change the representation into numerical value with respect to their category.

Table 3.3 Variables considered in the analysis by Deep Neural Network (DNN) and their values for six example patients.

Numb	Variable name	No.1	No. 2	No. 3	No. 4	No. 5	No. 6
1.	Diagnosis Sub-Type	1	2	2	3	3	3
2.	Diagnosis Type	1	4	1	1	1	3
3.	Department	19	6	2	2	2	1
4.	Government	5	5	1	6	2	9
5.	patient jobs	1	1	2	2	2	2
6.	blood group	1	2	2	2	2	3
7.	Gender	2	1	2	1	1	2
8.	department name	1	2	2	3	3	3
9.	Type	In	out	out	In	out	In

b. Data Transformation

The transformation technique is used to transform the values data set into numerical numbers. We have also applied this technique to our data set to convert all the strings into natural numbers by using find-replace.

In fact, the aim of data discretization to discretize attributes of continuous as this a method, which receives a data set, and changes completely attributes of continuous into categorical. This technique has been carried out by allocating every value in a

data set. In our data set as well, continuous discretized by using data discretization technique.

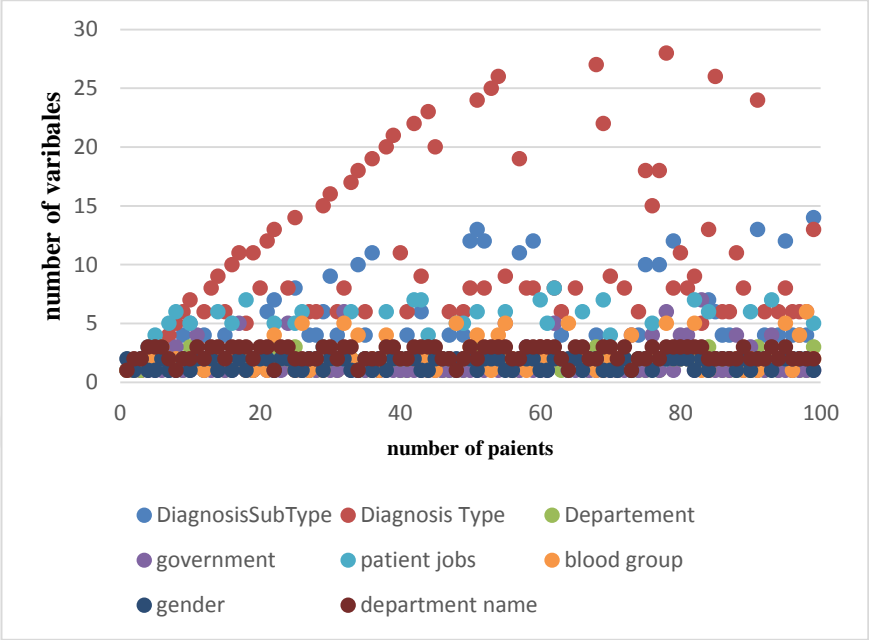


Figure 3.2 Shows the projection of 100 points denoting patients these points for 8 variables listed in the previous table (100 *8) data point

c. Data Normalization

This technique is a process to decrease data to its Standard form. In data normalization technique, attributes normalized by measuring their values; they drop into a small-distinguished range. The below formula used to normalize the class attribute:

$$v = \frac{v - min_A}{max_A - min_A}(new.max_A - new.min_A) + new.min_A \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.1)$$

Table 3.4 Data normalization

Variable number	Variable name	No.1	No. 2	No. 3	No. 4	No. 5	No. 6
1.	Diagnosis Sub-Type	0.01351	0.02702	0.03374	0.01351	0.02702	0.0135
2.	Diagnosis Type	0.02890	0.034682	0.04046	0.01734	0.03468	0.0462
3.	Department	0.6666	0.333333	1	0.66666	0.33333	0.666
4.	Government	0.125	0.041666	0.04166	0.16666	0.04166	0.041
5.	Patient Jobs	0.15	0.05	0.125	0.05	0.05	0.05
6.	Blood Group	0.2222	0.2222	0.22222	0.22222	0.111	0.222
7.	Gender	0.33333	0.6666	0.33333	0.66666	0.666	0.666
8.	Department Name	0.2	0.4	0.4	0.6	0.4	0.6
9.	Type	In patient	out patient	out patient	In patient	out patient	In patient

CHAPTER 4

RESULT AND DISCUSSION

4.1 INTRODUCTION

In this chapter, the main purpose is to evaluate the proposed algorithm for classifying cancers with respect to the medical record datasets using deep neural networks (DNN) to classify the instances based on their categorizes(in patients, out patients). This chapter presents experimental results for the proposed algorithm to evaluate the effectiveness of deep neural network algorithm for classification problem. The proposed algorithm intends to classify the data comparing with the basic artificial neural network. Section 4.3 presents the experimental setting discussed. Section 4.4, explains the results for DNN versus ANN algorithms. Section 4.5 discusses experiment results obtained by DNN algorithm. Section 4.6 concludes the results obtained.

4.2 EXPERIMENTS SETTING

There are several experiments that have been performed empirically to evaluate different feature sizes, training and testing data set. The experiment investigates the effectiveness of proposing deep neural network (DNN) to classify cancers to specific diagnostic categories based on their medical record.

These cancers belong to two distinct diagnostic categories (Inpatient and Outpatient) and often present diagnostic dilemmas in clinical practice. The performance evaluation of the proposed algorithm will be compared versus the state of the art algorithms in order to achieve the research objectives. The dataset has two different labels to categorize the patient type into two categories, which are Inpatient and Outpatient. Using different training and testing datasets evaluate the deep neural network (DNN) and artificial neural network (ANN). The training dataset are used to train the DNN and ANN classifiers. The evaluation metrics used are the

Classification rate and root mean square error RMSE of training data and testing data using different features sizes.

4.3 Experiment results for DNN against ANN algorithms

In the experiment, DNN and ANN algorithms have been implemented for their comparison with each other's on different training and test set using 70:30, 80:20 and 60:40 respectively. Table 4 presents the classification accuracy for the training and testing datasets that have used in each experimental run.

As it is clearly seen from the experimental results, the performance of the DNN algorithm is superior to ANN model. The best performance was scored when using 70:30 training and testing data sizes, the scored values were 73.97% and 66.6% for DNN and ANN respectively. The DNN achieves better results due to the amount of information that allows the classifier to increase the learning process and achieve a high accuracy. In the other side, the worst performance scored when using 60:40 for training and testing data sizes. The obtained results were (71.73 and 64.4) for DNN and ANN respectively.

Table 4.1 Results for DNN with respect to the different size of each set of data.

M	Size of individual sets of data (learning: testing)	DNN Classification Accuracy	ANN Classification Accuracy
1	80:20	73.63	63.1
2	70:30	73.97	66.6
3	60:40	71.73	65.4

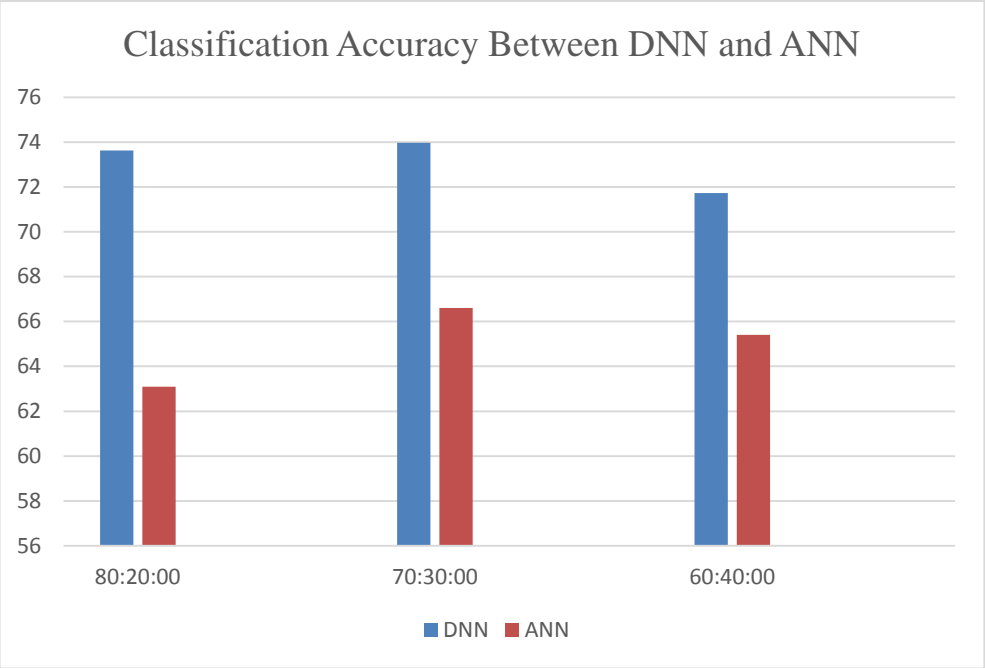


Figure 4.1 shows the experimental results for DNN and ANN algorithms

4.4 Results for Deep Neural Network (DNN)

The neural network has been trained using a different sample of input data, an example of data structure with their values for the selected features. The most significant key of training is to enhance the ability of the network to classify the data. These training data are strongly depends on the training size that used in the experiment. In addition, the data should contain a sufficient number of information to allow the network to learn by extracting the hidden structure in the dataset and then use this “knowledge” to extract a rule to new cases.

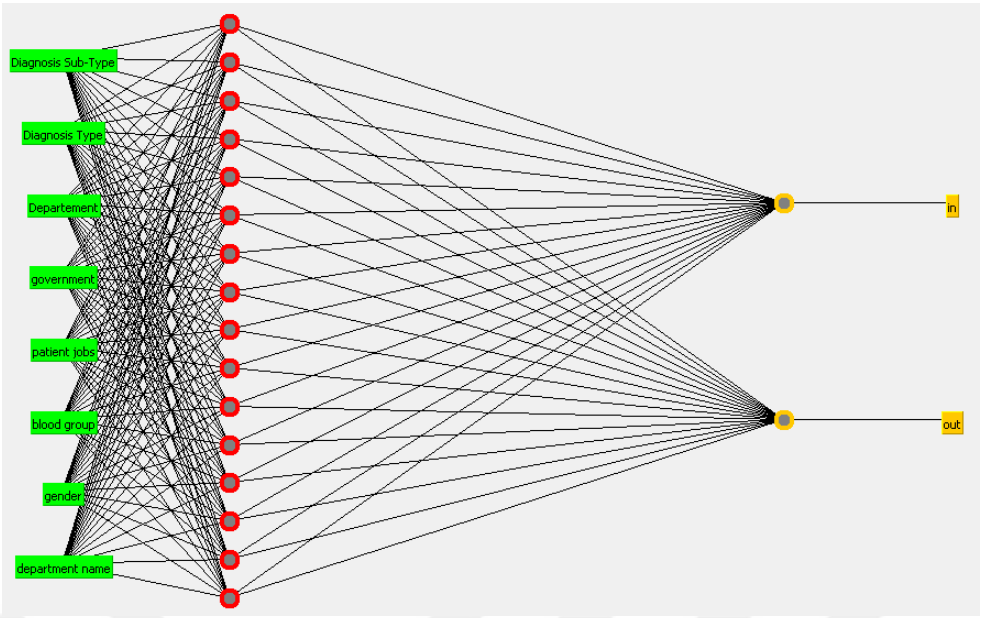


Figure 4.2 The architecture of deep neural network used in predictions of Iraq hospital dataset.

DNN consists of eight neurons in the input layer, eight in the hidden layer and one neuron in the output layer. In order to understand the DNN architecture, the figure shows the variables for patients analyzed are divided into two sets: learning set and testing set.

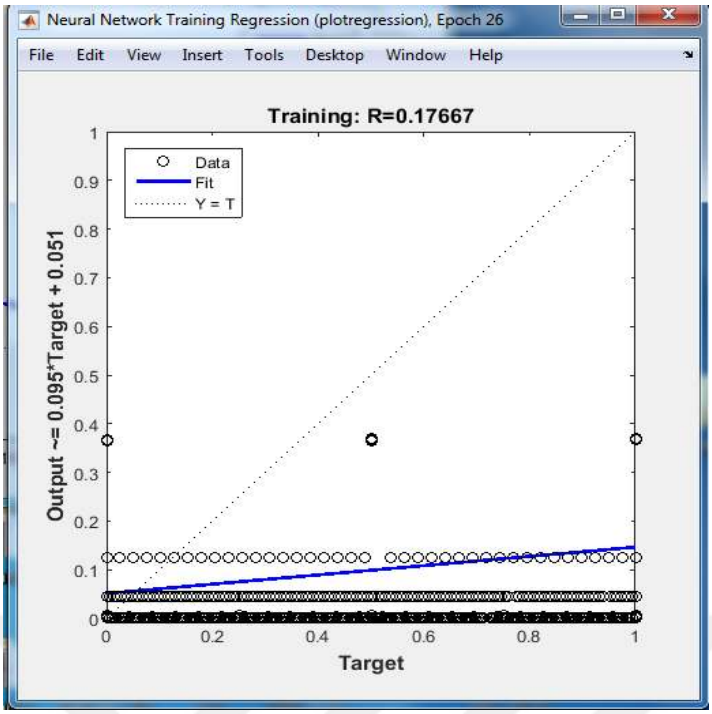


Figure 4.3 shows the training dataset

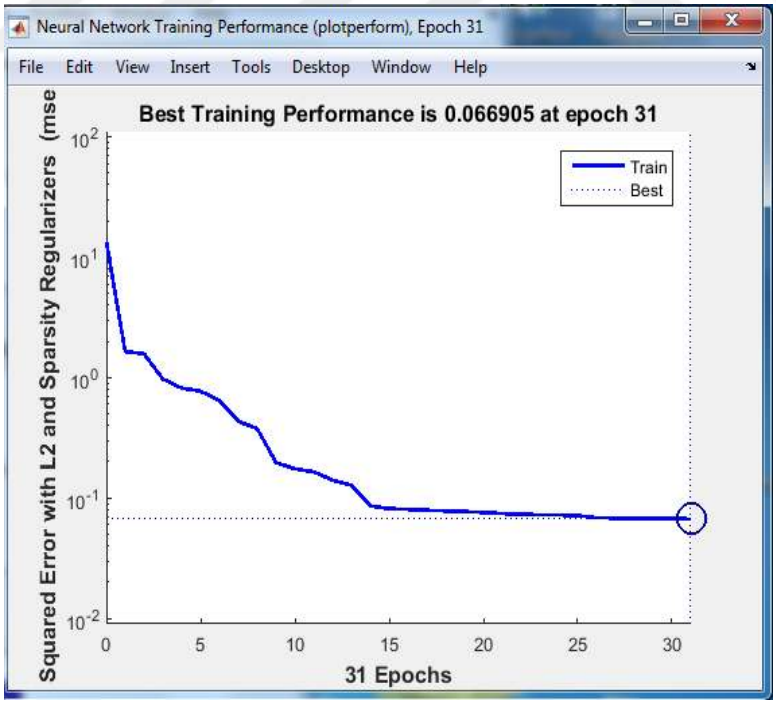


Figure 4.4 shows the best training

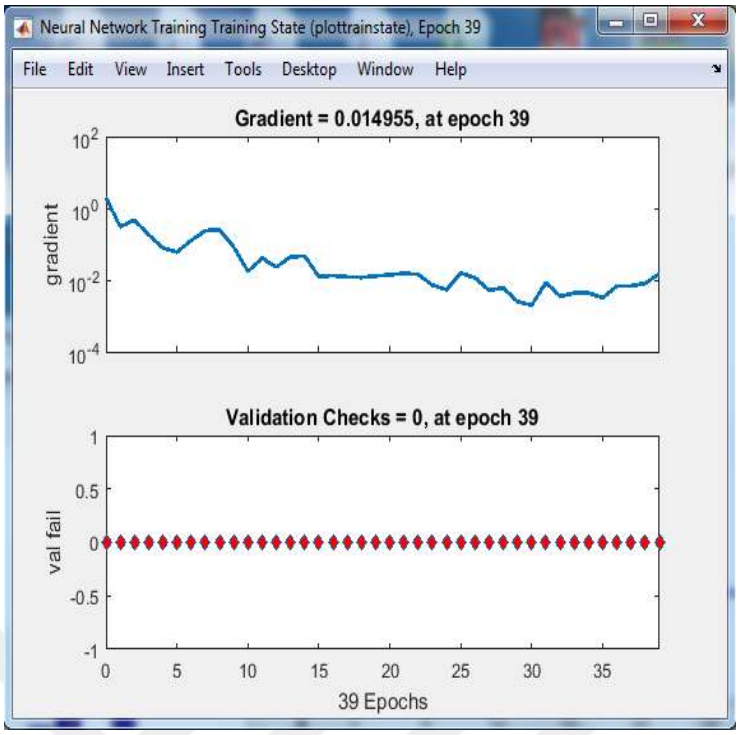


Figure 4.5 Validate the data at epoch 39.

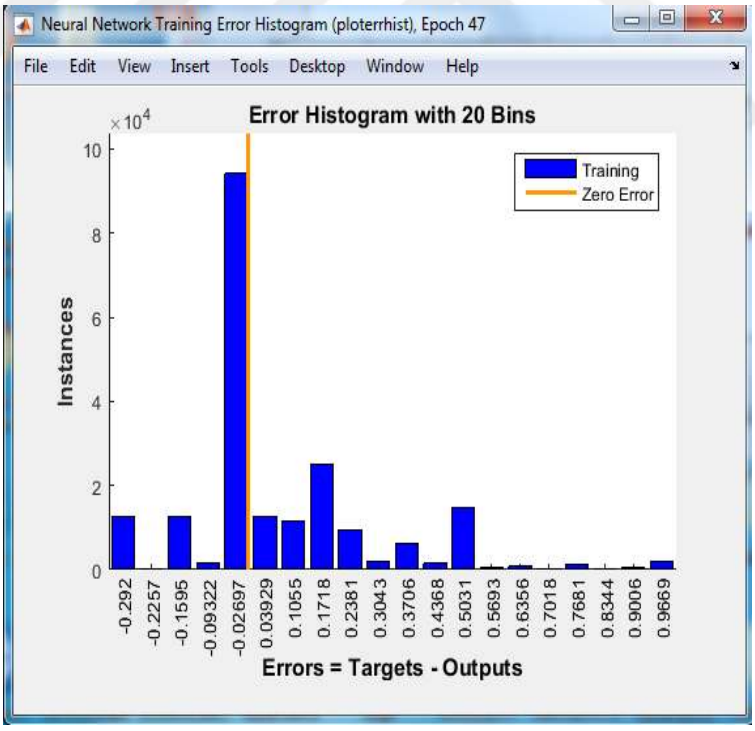


Figure 4.6 Histogram error with different bin

According to [24] the learning rate is the most important hyper-parameter to tune. Using a too large learning rate may cause the training algorithm to have problem

with convergence, and a too small learning rate may get the algorithm to get stuck in a local minima with bad generalization. For this study an adaptive learning rate, large in the beginning and smaller at the end, have been used with the purpose of finding good minima and reaching the optimum at that minima.

.

The learning process completed when the deep neural network was characterized by the least RMS error. In the case of the network applied, learning was completed in 500 epochs by DNN method. In this stage, deep neural network used two hidden layer, momentum 0.2 and learning rate 0.3 in all tests. Data from the learning set was presented in a randomized manner during the learning process.

Table 4.2 Results for DNN with respect to the different size of each set of data.

Model	Size of individual sets of data (learning: testing)	DNN model type	Classification rate of training data	Classification rate of Testing data
1	80:20 20958:5240	DNN 8:8-16-8-1: 1	62.84	73.63
2		DNN 8:8-8-8-1: 1	62.60	72.84
3		DNN 8:8-12-4-1: 1	61.60	68.36
4	70:30 18338:6312	**DNN 8:8-16-8-1: 1**	**63.41**	**73.97**
5		DNN 8:8-8-8-1: 1	61.38	73.45
6		DNN 8:8-12-4-1: 1	61.51	73.11
7	60:40 15718:10480	DNN 8:8-16-8-1: 1	62.46	71.73
8		DNN 8:8-8-8-1: 1	60.60	71.59
9		DNN 8:8-12-4-1: 1	60.66	73.05

Table 4.2 presents the results that have been scored using different training and testing datasets (80:20, 70:30 and 60:40). In addition, we have used different DNN models for DNN to test the performance with a different model types. The obtained results indicated that, the best classification rate is scored reaches up to 73.64 when using 70:30 training and testing and DNN 8:8-16-8-1: 1 model.

Table 4.3 Classification results for six DNN models with a different number of hidden neurons.

Model	Type of DNN	Classification rate of training data	Classification rate of testing data
1	DNN 8:8-6-3-1: 1	**84.56**	**87.84**

2	DNN 8:8-9-10-1: 1	62.84	73.63
3	DNN 8:8-15-1: 1	77.18	78.85
4	DNN 8:8-16-1: 1	82.92	81.47
5	DNN 8:8-4-1: 1	73.76	82.90
6	DNN 8:8-5-5-1: 1	67.97	76.91

According to the results in Table 4.3, we summarize that the best accuracy scored when using model (1, 4 and 5) the classification rate scored was 87.84, 81.47 and 82.90 respectively. The best performances are scored when using model 1 which archives an accuracy of 87.84.

Moreover, the performance of the other model achieves lower results compared to the best models; the performance of the model 2, 3 and 6 was below 78.58. The results demonstrated the effect of having a large number features on the classifier results. Therefore, the results performance seems to be similar due to the large number of feature sizes. Enhancement algorithms for feature selection should be acknowledged for further investigation.

Table 4.4 Classification result for DNN model using 70:30 training and testing dataset.

Model	Type of DNN	Learning set		Testing set	
		In Patient	Out patient	In Patient	Out patient
1	Total	7083	11255	2087	5773
2	Correct	5448	9395	1862	4450
3	Wrong	1635	1860	225	1323

Based on the obtained results, we should shed a light on the results that obtained when using 70:30 for training and testing. Table 4.4, presents the dataset size for 70:30 dataset. The above numbers indicates the number of total instances for training and testing for inpatient and outpatient.

For learning set, the total number of in-patient and out-patient was 7083 and 11255 respectively. The total number of correctly classified was 5448 and 9395 instances; the wrongly classified was 1635 and 1860 instances.

Table 4.5 Sensitivity analysis results for the variables considered in deep neural networks DNN analysis.

Model	Type of DNN	RMSE Training	RMSE Testing data
1.	DNN 8:8-6-3-1: 1	0.4008	0.3975
2.	DNN 8:8-15-1: 1	0.4539	0.4411
3.	DNN 8:8-16-1: 1	0.4712	0.4759
4.	DNN 8:8-4-1: 1	0.4066	0.3524
5.	DNN 8:8-5-5-1: 1	0.42245	0.3884
6.	DNN 8:8-16-8-1: 1	0.6096	0.5136
7.	DNN 8:8-8-8-1: 1	0.4733	0.4492
8.	DNN 8:8-12-4-1: 1	0.4596	0.4367
9.	DNN 8:8-16-8-1: 1	0.60	0.51
10.	DNN 8:8-8-8-1: 1	0.6215	0.5153
11.	DNN 8:8-12-4-1: 1	0.6215	0.5153
12.	DNN 8:8-16-8-1: 1	0.4661	0.4487
13.	DNN 8:8-8-8-1: 1	0.6277	0.5330
14.	DNN 8:8-12-4-1: 1	0.4890	0.4565

Table 4.5 lists the sensitivity analysis results for the DNN after identifying the best training and testing set. The performance evaluation was based on different model sizes to evaluate the RMSE for training and testing datasets.

The obtained results demonstrate that the best performance scored when using model number 4 with DNN 8:8-4-1: 1. The RMSE value scored using 0.4066 for training and 0.3524 for the testing dataset. In addition, the worst results scored when using model 13 the performance of the model was scored 0.6277 for training and 0.5330 for testing.

On the other hand, the total number of testing set for in-patient and out-patient was 2087 and 5773 instances. The total number of the correctly classified was 1862 and 4450 instances and the wrongly classified was 225 and 1323 instances.

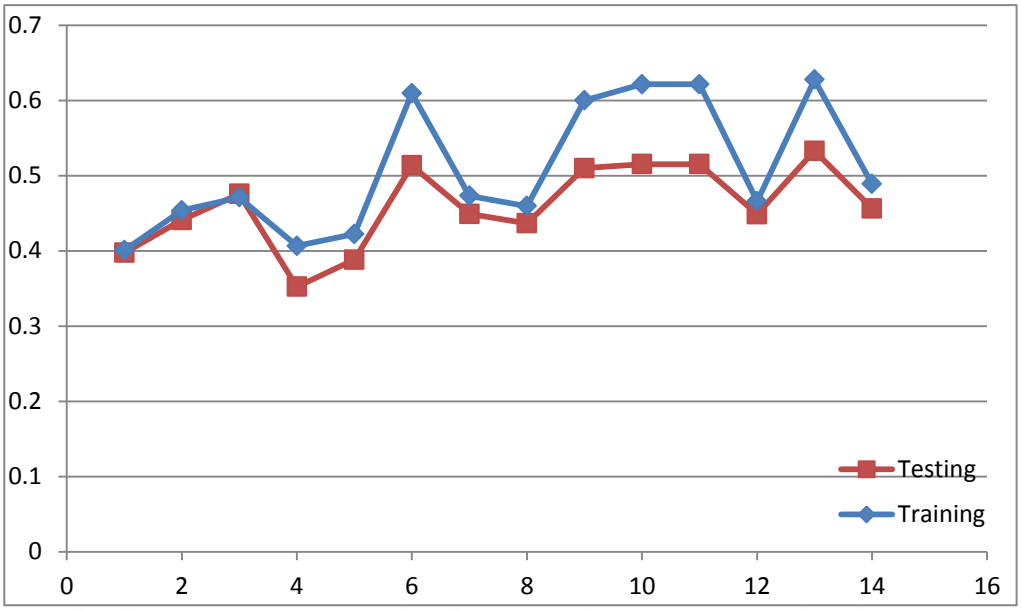


Figure 4.7 shows the RMSE for the training and testing using different DNN model.

Figure 4.7 Depicted the Root mean square error for the training and testing dataset with different DNN models. Based on the depicted figure above, it's clearly seen that the error percentage are lower than the training data set, which is indicated that the DNN in the training dataset manages to find a pattern that decreased the error percentage in the testing data.

4.5 Result Analysis

Deep neural networks (DNN) have been implemented on computer simulation platform. At first, the experiment has conducted to determine the best size of individual sets of data (learning: testing) using three different data size for training and testing which are (80:20, 70:30 and 60:40). Table 8 presented the data sizes along with the performance for training and testing for each data size. In addition, different DNN model is tested to determine the best model.

based on the obtained results, the performance of the 70:30 dataset has achieved higher accuracy compared with the other DNN models, the best results scored when using DNN 8:8-16-8-1: 1 model. The best classification rate reached to 63.41 and 73.97 for training and testing respectively. For further investigation, a deep neural network based on six DNN models with a different number of hidden neurons was used. The training and testing dataset 70:30 were used to test the performance. The

best Classification rate of training and testing data were recorded when using DNN 8:8-6-3-1: 1 model. The result reaches up to 84.56 and 87.84 respectively.

CHAPTER 5

CONCLUSION & RECOMMENDATION

5.1 INTRODUCTION

In this study, we suggested the deep neural network algorithm in order to classify the cancers to specific diagnostic or categories based on their medical record using deep neural networks (DNN).

In this work, we have investigated the performance of DNN algorithm and compared the obtained results with the classic ANN algorithm, which aims to extract a pattern from healthcare data.

5.2 Conclusion

In conclusion, the main objective of this work is to classify the performance of health patient dataset. The deep neural network (DNN) is implemented and the obtained results are compared against the artificial neural network ANN. The results analysis shows that the methodology is adequate and efficient for classifying the patient type Also In this thesis, we propose a method to enhance cancer diagnosis and classification from expression Data using deep neural network methods Applying this method to cancer data and comparing it to baseline algorithms, our method not only shows that it can be used to improve the accuracy in cancer classification problems, but also demonstrates that it provides a more general and scalable approach to deal with gene expression data across different cancer types.

This thesis introduced the development of KDD and the overall system, and then proposes a deep neural network for feature extraction and the Computer simulation program classifier for feature extraction. Finally, the different training strategies are discussed

5.3 Future Work

For future works, there are several issues that require more exploration and need to be investigated by the research community for healthcare data in order to classify the cancer based on their categorize These issues, which are discussed in the following points, have been identified based on the obtained results in section 4.

- Use feature selection algorithms are required in order to enhance the performance of the selected features such as ant colony algorithm or particular swarm algorithm.
- Propose a novel algorithm that has the ability to automatically assign the proper parameters to enhance the classification accuracy.

REFERENCES

[1] V. Chandola, S. R. Sukumar, and J. C. Schryver, "Knowledge discovery from massive healthcare claims data," in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013, pp. 1312-1320.

[2] A. Holzinger and I. Jurisica, "Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions," in Interactive knowledge discovery and data mining in biomedical informatics, ed: Springer, 2014, pp. 1-18.

[3] G. Piateski and W. Frawley, *Knowledge discovery in databases*: MIT press, 1991.

[4] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine,* vol. **17**, p. 37, 1996.

[5] Y. Wang, R. Chen, J. Ghosh, J. C. Denny, A. Kho, Y. Chen*, et al.*, "Rubik: Knowledge guided tensor factorization and completion for health data analytics," in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 1265-1274.

[6] N. Esfandiari, M. R. Babavalian, A.-M. E. Moghadam, and V. K. Tabar, "Knowledge discovery in medicine: Current issue and future trend," *Expert Systems with Applications,* vol. 41, pp. 4434-4463, 2014.

[7] H. Q. Yu, X. Zhao, Z. Deng, and F. Dong, "Ontology driven personal health knowledge discovery," in *International Conference on Knowledge Management in Organizations*, 2015, pp. 649-663.

[8] C. Cortes and V. Vapnik, "Support vector machine," *Machine learning,* vol. 20, pp. 273-297, 1995.

[9] H. R. Marucci-Wellman, M. R. Lehto, and H. L. Corns, "A practical tool for public health surveillance: Semi-automated coding of short injury narratives

from large administrative databases using Naïve Bayes algorithms," *Accident Analysis & Prevention,* vol. 84, pp. 165-176, 2015.

[10] P. Miasnikof, V. Giannakeas, M. Gomes, L. Aleksandrowicz, A. Y. Shestopaloff, D. Alam*, et al.*, "Naive Bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths," *BMC medicine,* vol. 13, p. 286, 2015.

[11] X. Zhou, S. Wang, W. Xu, G. Ji, P. Phillips, P. Sun*, et al.*, "Detection of Pathological Brain in MRI Scanning Based on Wavelet-Entropy and Naive Bayes Classifier," in *IWBBIO (1)*, 2015, pp. 201-209.

[12] L. E. Peterson, "K-nearest neighbor," *Scholarpedia,* vol. 4, p. 1883, 2009.

[13] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Advances in neural information processing systems*, 2006, pp. 1473-1480.

[14] M. M. Adankon and M. Cheriet, "Support vector machine," *Encyclopedia of biometrics,* pp. 1504-1511, 2015.

[15] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector machine learning for interdependent and structured output spaces," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 104.

[16] W. G. Baxt and J. Skora, "Prospective validation of artificial neural network trained to identify acute myocardial infarction," *The Lancet,* vol. 347, pp. 12-15, 1996.

[17] M. H. Ebell, "Artificial neural networks for predicting failure to survive following in-hospital cardiopulmonary resuscitation," *Journal of family practice,* vol. 36, pp. 297-304, 1993.

[18] I. Aleksander and H. Morton, *An introduction to neural computing* vol. 3: Chapman & Hall London, 1990.

[19] S. S. Cross, R. F. Harrison, and R. L. Kennedy, "Introduction to neural networks," *The Lancet,* vol. 346, pp. 1075-1079, 1995.

[20] V. Kecman, *Learning and soft computing: support vector machines, neural networks, and fuzzy logic models*: MIT press, 2001.

[21] D. Yu, L. Deng, F. T. B. Seide, and G. Li, "Discriminative pretraining of deep neural networks," ed: Google Patents, 2016.

[22] Y. Zhou, X. Zheng, and W. Wu, "Effect of local information within network layers on the evolution of cooperation in duplex public goods games," *Chaos, Solitons & Fractals,* vol. 78, pp. 47-60, 2015.

[23] Y. Bengio, I. J. Goodfellow, and A. Courville, "Deep learning," *Nature,* vol. 521, pp. 436-444, 2015.

[24] H. Chabanne, A. de Wargny, J. Milgram, C. Morel, and E. Prouff, "Privacy-Preserving Classification on Deep Neural Network," *IACR Cryptology ePrint Archive,* vol. 2017, p. 35, 2017.

[25**]** Paolo G. "Applied Data Mining Statistical Methods for Business and Industry", John Wiley & Sons Ltd, Chichester, England, 2003.

[26] Bramer M. "Principles of Data Mining", British Library Cataloguing in Publication Data, Springer Science+Business Media -Verlag London, 2007.

[27] Sirikulvadhana S., "Data Mining As A Financial Auditing Tool", M.Sc. Thesis, Dept. of Accounting, Swedish School of Economics
And Business Administration, 2002.

[28] Wang A. "Data Mining and Statistics: Examining Critical Patterns of Research and Practi*ce"*, Ph.D. Thesis, Dept. of Statistics, University of Texas Arlington, May 2006.

[29] Bandyopadhyay S., Maulik U., Holder B., and Cook J. "Advanced Methods for Knowledge Discovery from Complex Data", Springer Science+Business Media, 2005.

[30] Cios J., Pedrycz. W., Swiniarski W. and Kurgan A. "Data Mining: AKnowledge Discovery Approach", Springer Science+Business Media, LLC, 2007

[31] Imberman P., "Comparative Statistical analyses of Automated Booleanization Methods for Data Mining Programs", Ph.D. Thesis, Dept. of Computer Science, University of New York, 1999.

[32] Sumathi S. and Sivanandam N. "*Introduction to Data Mining and its Applications*", Springer Science+Business Media, Springer- Verlag Berlin Heidelberg, 2006.

[33] Liu B. "*Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*", Springer Science+Business Media, Springer-Verlag Berlin Heidelberg, 2007.

[34] Hand D., Mannila H. and Smyth P. "*Principles of Data Mining*",The MIT Press,ISBN: 026208290x, 2001.

[35] Matthias. S., "*Advanced Data Mining Techniques for Compound Objects*", Ph.D. Thesis, Dept. of Informatics and Statistics, University of Munchen, (November 2004).

[36] Sirikulvadhana S., "*Data Mining As A Financial Auditing Tool"*,M.Sc. Thesis, Dept. of Accounting, Swedish School of Economics and Business Administration, 2002.

[37] Olson L. and Delen. D. "*Advanced Data Mining Techniques*",Springer-Verlag Berlin Heidelberg, 2008.

[38] Larose T. "*Data Mining Methods and Models*", John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.