KADIR HAS UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

USE OF MACHINE LEARNING TECHNIQUES FOR DIAGNOSIS OF THYROID GLAND DISORDER

IZDIHAR MOFEK

May 2016

USE OF MACHINE LEARNING TECHNIQUES FOR DIAGNOSIS OF THYROID GLAND DISORDER

IZDIHAR MOFEK

Submitted to the Graduate School of Science and Engineering

In partial fulfillment of the requirements for the degree of

Master of Science

In

Computer Engineering

KADIR HAS UNIVERSITY

May, 2016

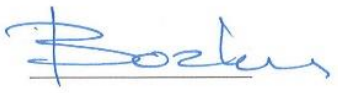
KADIR HAS UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

USE OF MACHINE LEARNING TECHNIQUES FOR DIAGNOSIS OF THYROID GLAND DISORDER

IZDIHAR MOFEK

APPROVED BY:

Assoc. Prof. Dr. Zeki BOZKUŞ (Thesis Supervisor)	Kadir Has University	
Asst. Prof. Dr. Tamer Dağ	Kadir Has University	
Assoc. Prof. Dr. Ahmet Yücekaya	Kadir Has University	

APPROVAL DATE: 23/05/2016

USE OF MACHINE LEARNING TECHNIQUES FOR DIAGNOSIS OF THYROID GLAND DISORDER

Abstract

The advancements of computer technologies have generated an incredible amount of data and information from numerous sources. Nowadays, the way of implementing health care are being changing by utilizing the benefits of advancements in computer technologies. It is believed that engineering this amount of data can assist in developing predictive tool that can help physicians to diagnosing and predicting some debilitating life-threatening illness such as thyroid gland disease.

Our current work focuses on investigating python languages to diagnose thyroid gland disease based on machine learning, and involves developing a new tool to predict the diagnoses of thyroid gland diseases, which we have called as a MLTDD (Machine Learning App for thyroid Disease Diagnosis). MLTDD has been designed with Qt designer and programmed using PyDev, which is python IDE for Eclipse. MLTDD could diagnose with 99.81% accuracy.

Decision tree algorithm has been used to create the ML model, in addition to training dataset to learn from. ML model can be used to get predictions on new data for which you do not know the target and that is what we did to predict the diagnosis of thyroid gland disease as a hyperthyroidism or hypothyroidism or a normal condition using CRT decision tree algorithm.

MLTDD can minify the cost, the waiting time, and help physicians for more research, as well as decrease the errors and mistakes that can be made by humans on account of exhaustion and tiredness.

Keywords: Machine Learning, Thyroid diseases, decision tree algorithm, PyDev, python IDE, ML model.

TİROİT BEZİ BOZUKLUĞU TANISI İÇİN MAKİNA ÖĞRENME KULLANIMI

Özet

Bilgisayar teknolojilerinin gelişmeler sayesinde sayısız kaynaklardan inanılmaz büyüklükte veriler yarattı. Bu verilerin işlenmesi, ve bilgiye dönüşümü yine bilgisayar teknolojileri tarafından yapılmakta. Günümüzde, sağlık uygulamalarında bilgisayar teknolojilerindeki gelişmeler faydalanıyor. Makine öğrenmesi araçlarını kullanarak verilerin daha kullanışlı haline getirilebilir. Tiroit bezi hastalığı gibi bazı hastalıkların teşhisinde doktorlara yardımcı olmak amacıyla, makina öğrenme algoritmalarının kullanılmaya başlanmıştır. Bu teknolojiler insanlar tarafından yapılabilir hataları ve yanlışlıkları azaltabilir.

Tiroit bezi tiroit hormonları kontrol altında tutmak ve metabolizmanın çalışma hızını ayarlayan, insan vücudu üzerinde önemli organlarından biridir. Tiroit vücudumuzdaki tüm organları ve hücreleri etkiler. Tiroit bezi çok fazla hormon üretir ise, vücut olması gerekenden daha hızlı enerji kullanır ve bu durum hipertiroidi olarak bilinir. Tiroit kendi hormon yeterli miktarda üretilmez ise, vücut olması gerekenden daha yavaş davranır ve az enerji kullanır ve bu duruma hipotiroidi denir. Tiroit hastalığı kolayca diğer hastalık koşulları ile karıştırılır, tiroit hastalığı teşhis etmek zor olabilir.

Etkili makine öğrenme algoritmaları tiroit bezi bozukluğu gibi hastalıkların erken teşhis için yardımcı olabilir. Bu tezde mevcut makinası öğrenme araçları dayalı tiroit bezi hastalığı teşhis etmek için Python dilinde kullanarak MLTDD isminde bir araç geliştirdik. MLTDD tiroit hastalığını etkili bir şekilde teşhisinde yardımcı olabilir. MLTDD karar ağacı algoritması kullanarak model oluşturup, yeni gelen hastaları bu algoritma sayesinde tiroit hastası olup olmadığını sınıflandırıyor. Ara yüzü, Qt tasarımcı ile tasarlanmış ve Eclipse için Python IDE Pydev kullanılarak programlanmıştır. MLTDD% 99.81 doğruluk ile teşhis yapabilmektedir.

Anahtar Kelimeler: Makine Öğrenmesi, Tiroit hastalıkları, karar ağacı algoritması, Pydev, Python IDE, ML modeli.

Acknowledgements

I would like to thank everyone has helped me to complete this thesis. Special, honest and deep thanks to my supervisor Assoc. Prof. Zeki Bozkuş for the guidance and his advice through my research. My big thanks to all faculty staff and employees, and I am pleased for being a small part of such great place for a couple of years, principally Prof. Dr. Feza Kerestecioğlu for his support provided through my journey.

My distinctive thanks to my husband; you were a total package of family and love that gave me strength, support during tough times. My daughters Weaam and Bayan, you are my words, my world, my strength, and the reason of my life. Thank you. Words can never express my thanks for my family, the source of power and strength; my parents for raising me up. My Father, thank you for everything you perfectly done to me, for the advice, encouragement and support through my life. My Mother, thanks are not enough and would not be enough, and sorry for being away from your warm and benign bosom but I will be back, hopefully soon. Thank you for my brothers and sisters.

I would like to tank Elaine Woods; you have been so much more than a teacher for me. Before meeting you, I never thought that a person could be a teacher, support, mentor and sage all in one. I am grateful to have been taught by such an admirable person. Thank you for everything you have done for me.

I would like to deeply thank my friend Ebru ALP for being such a helpful friend, and I am grateful to my friend Fatma Al-Nemsy for her nice collaboration.

Thank you all, I hope that in the future we can continue our fruitful cooperation.

Table of Contents

Abstract i

Özet.....……...……..…..ii

Acknowledgements iii

Table of Contents..... iv

List of Figures..... vii

List of Tables..... ix

List of Abbreviations..... x

Chapter 1: Introduction 1

1.1 Overview..... 1

1.2 Thesis Road Map..... 2

Chapter 2: Motivation and Research Objectives..... 4

2.1 Research Motivation.....4

2.2 Research Objectives..... 8

2.3 Research Contribution..... 10

Chapter 3: Background Study and Literature Review..... 12

3.1 Background Study..... 12

3.2 Machine learning..... 12

3.2.1 Snapshot Applications of Machine Learning…..…..10

3.3 Machine learning in Healthcare…….……....…..……15

3.4 Machine Learning Algorithms.....16

3.4.1 Algorithms Grouped by Learning Style.....16

3.4.2 Algorithms Grouped By Similarity 18

3.5 Classification 17

Chapter 4: Research methodology.....22

4.1 Thyroid gland disorder.....22

4.1.1 Thyroid gland system.....22

4.1.2 Effects of Thyroid Hormones.....24

4.1.3 Thyroid Diseases.....25

4.2 Description of data set.....27

4.2.1 General Description.....27

4.2.2 Attribute Description.....29

4.2.3 Data set visualization.....32

4.3 Decision Tree algorithm.....33

4.3.1 Decision tree representation.....33

4.3.2 Decision tree types.....34

4.3.3 Decision tree Metrics.....35

4.3.4 Popular Decision Tree Classification & Regression Trees.....36

Chapter 5: Developing a ML tool for Thyroid Disease Diagnosis.....39

5.1 Machine learning in Python.....39

5.1.1 Python programming language.....39

5.1.2 Python IDE (Integrated development environment).....40

5.2 Applying CART algorithm.....41

5.2.1 Algorithm fitting.....41

5.2.2 Evaluation.....45

5.2.3 Tree visualization.....48

5.3 Designing the Graphical user interface (GUI) 49

5.3.1 PyQt based application and QT designer……..……50

5.3.2 Initial design.....51

5.3.3 After pre-processing……..……..……..…51

5.4 Using Eclipse (PyDev) to connect GUI with python files.....53

5.5 Performance Results …….....57

Chapter 6: Discussion and Related work.....60

Conclusion.....….64

References.....……65

Curriculum Vitae ……..…….....…...69

List of Figures

Figure 1: Estimated Thyroid gland incidence and mortality rate……..…6

Figure 2: Estimated Thyroid incidence and mortality rate as percentage…6

Figure 3: Prevalence rate for Thyroid disorders of different countries 7

Figure 4: Research Method overview ……..…..…...…...9

Figure 5: Supervised Learning algorithms …….....…17

Figure 6: Unsupervised Learning algorithms ……...…….....…18

Figure 7: General approach for building a classification model ……20

Figure 8: Thyroid system …….....…24

Figure 9: Python code to get the number of instances per class ……29

Figure 10: Python code to get the domain of each attribute ……..…32

Figure 11: Class visualization for the whole dataset …….……32

Figure 12: Importing the required Python libraries……..…..41

Figure 13: Converting the dataset into a pandas frame …….42

Figure 14: Sample of Ann-Thyroid Diagnosis dataset……..…..……42

Figure 15: Histogram for Ann_tyroid dataset……..…..……43

Figure 16: Predicting the class label for training and testing dataset.……45

Figure 17: Classification and prediction accuracy……..……46

Figure 18: Precision and recall for class label…….……46

Figure 19: Classification report for training dataset ……..…..……47

Figure 20: decision tree visualization……..…….…….…..49

Figure 21: Snapshot of the initial design of MLTDD……..…51

Figure 22: Information Gain Ranking Filter……52

Figure 23: Snapshot MLTDD interface……...……53

Figure 24: Converting gender data to numeric data ……...…..…55

Figure 25: Converting the age data …….……55

Figure 26: Apply and train CART algorithm …….…..……56

Figure 27: Setting button actions ……56

Figure 28: Accuracy comparison graph for the splitting percentage ..……59

List of Tables

Table 1: The confusion matrix for classifier ……. 21

Table 2: Thyroid Disease (ann) data set General information ……..….…28

Table 3: Thyroid Disease (ann) data set Attribute description ……..…31

Table 4: Accuracy comparison for three algorithms.....…..58

Table 5: Accuracy before and after eliminating Ten features……..…59

List of Abbreviations

EHR	Electronic Health Record
NWHIC	National Women's Health Information Center
WHO	World Health Organization
UK	United Kingdom
US	United States
GUI	Graphical User Interface
ML	Machine Learning
UCI	University of California of Iravin
CRT	Classification and Regression Trees
WYSIWYG	What-You-See-Is-What-You-Get
PyDev	Python Development
AI	Artificial Intelligence
NLP	Natural Language Processing
PCA	Principal Component Analysis
kNN	k-Nearest Neighbour
ID3	Iterative Dichotomiser 3
LDA	Linear Discriminant Analysis
MDA	Mixture Discriminant Analysis
QDA	Quadratic Discriminant Analysis
FDA	Flexible Discriminant Analysis
TRH	Thyrotropin Releasing Hormone

TSH	Thyroid Stimulating Hormone
FT4	Free T4
TDIDT	Top-Down Induction of Decision Trees
CHAID	CHi-squared Automatic Interaction Detector
IDE	Integrated Development Environment
NumPy	Numerical Python
SciPy	Scientific Python
SDK	Software Development Kit
CAS	Computer Algebra System
IP[y]	Interactive Python
ANNs	Artificial Neural Networks
LVQ	Learning Vector Quantization
RBFN	Networks, the Radial Basis Function
BPA	Back propagation algorithm
CSFNN	Conic Section Function Neural Network
FBP	Fast Back-Propagation
MLP	Multi-Layer Perceptron
GDA-WSVM	Generalized Discriminant Analysis-Wavelet Support Vector Machine

Chapter 1

Introduction

1.1 Overview

The huge advancement of information technology, system integration techniques and software development have produced a new generation of complicated computer systems. Therefore, these systems have presented big challenges to computer engineering researchers. A good example of these complex systems is the healthcare system. Recently, it has been an increased interest to use the advancement of communication, data mining and machine learning technologies in healthcare systems. As a result, many countries are changing the way of organizing healthcare systems towards a global healthcare system across this country by setting healthcare standardization in communication and structuring(building) the electronic healthcare records.

The Electronic Health Record (EHR) is an organized collection of electronic health data about individual patients or some populations. It is qualified for sharing across healthcare providers in a certain country [1]. Health records normally contain a range of data including general medical records, medical history, patient examinations, laboratory results patient treatments, radiology images, allergies, immunization status, and some useful information for examination. Therefore, this valued information probably help researchers in examining and diagnosing diseases by using computer techniques. The use of EHRs may help in improving the quality of care, reducing the cost of legacy systems, and mobility of records.

Therefore, our aim in the current work is to investigate the aspects of utilizing the repository of health data for the benefit of humans by using machine learning. Our idea is to propose an automated method for diagnosing diseases based on previously stored

data and information. However, there are many problems related to effectively using this previously obtained patient data, which could make any electronic tool less efficient. Some of these problems are: the issue of huge features or attributes and how to select the most beneficial ones, the problem of missing values and how to process that, the problem of extracting accurate diagnostic markers that can predict the early start of the disease. This thesis will try to investigate some of these issues and propose predictive tool for thyroid gland disease diagnosis, based on the potential and the power of automated technologies and the previous patients or data. However, the scope of the thesis is exclusive to the problems outlined above, and does not include other equally important problems like privacy and security.

In this research, UCI repository will be used as our data sources for developing automatic machine learning tool, in order to produce useful predictive method for diagnosis thyroid gland disease. For Keeping track of the investigations for this project, the study used well-known datasets, which are publicly available for research purposes. It is planned, that the tool developed based on decision tree algorithm techniques, and validating on this dataset can be extended to real clinical environments.

1.2 Thesis Road Map

The thesis is organized into six chapters:

- Chapter 1: Introduction
- Chapter 2: Motivation and Research Objectives
- Chapter 3: Background Study and Literature Review
- Chapter 4: Research Methodology

- Chapter 5: Developing a ML tool for Thyroid Disease Diagnosis (MLTDD)

- Chapter 6: Discussion and Related Work.

After the introductory chapter (Chapter 1), which presents the problem description and the road map of thesis, Chapter 2 show the motivation and objectives of this work and the contribution to the scientific knowledge.

Chapter Three provides a review of machine learning techniques, and its application in the field of healthcare in addition to Machine Learning Algorithms. Chapter 4 describes the methodology used in this work. Chapter 5 presents the stages of designing a new app for thyroid gland diagnosis using machine learning techniques and the power of eclipse and python language. The idea is to obtain an intelligence app that combines the best performing learning algorithms and the powerful of GUI designing. Finally, Chapter 6 presents some of the conclusions generated from this work and scope for future work.

Chapter 2

Motivation and Research Objectives

2.1 Research motivation

Approximately one-third of the world's population lives in countries in areas of iodine deficiency [2]. Some areas where the daily iodine intake is less than < 50 μg so goiter is usually endemic, and when the daily intake of iodine falls under <25 μg congenital hypothyroidism is seen. The spread of goiter in areas of significant iodine deficiency can be as high as 80%. Iodization programs are of proven importance in reducing goiter size also in preventing goiter development and finally the cretinism in children. The iodization programs can also involve thyrotoxicosis, especially in those aged under >40 years with nodular goiters. In iodine-replete areas, most people with thyroid disorders often have autoimmune disease, ranging between primary atrophic hypothyroidism, to Hashimoto's thyroiditis to thyrotoxicosis which caused by Graves' disease [3].

Regarding to Goiter and thyroid nodules the most common thyroid disease in the community is common physiological goiter. In some surveys, the prevalence of diffuse goiter turns down with age; the highest prevalence is in pre-menopausal women thus the ratio of women to men is at least 4:1 [4]. This is in contrast to the increase in spread of thyroid antibodies and thyroid nodules with age. A study shows that5234 subjects aged more than 60 years in (Massachusetts), clinically apparent thyroid nodules were existing in 1.5% of men and 6.4% of women [5]. The prevalence of solitary thyroid nodules was 3% and multinodular goiter was 1%.

In many early necropsy surveys, more than 50% of patients had thyroid nodules and between 20 and 76% of women have at least one thyroid nodule [2]. In a relative iodine deficiency area in Germany, thyroid nodules or goitre were found in 33% of 96 278

working adults of age between 18 and 65 years [6]. Thyroid nodules bigger than 1 cm were found in 12% of this population and this percentage increased with age. Patients with a single sensible nodule, 20–48% have additional nodules.

The clinical display of thyroid cancer is generally as a solitary thyroid nodule or increasing goiter size. Thyroid cancer is the most common malignant endocrine tumor and it has considers to cause >90% of the cancers of the endocrine glands in the United Kingdom [7]. The incidence of thyroid cancer is growing. In 2001, data from Cancer Research UK displayed 1200 new cases in England and Wales, according to a reported yearly incidence for the UK of 3.5 per 100 000 women and 1.3 per 100 000 men [8].

In 2012 according to world health organization WHO report Estimated age-standardized incidence is 298102 approx. 2.1% and mortality 39771 approx. 0.5% .After 5 years of prelevanec that would be 1206075 approx. 3.7 %. Rates due to thyroid cancer for both sexes as demonstrated in figur1 as a thousands and figure 2 as percentage [9].

Congenital hypothyroidism affects nearly one newborn in 3500–4000 births and is the most treatable reason of mental retardation [10]. There is a converse relationship between age when diagnosing congenital hypothyroidism and intelligence quotient in later life. In iodine-replete areas, about 85% of the cases are due to interrupted developmental defects of the thyroid gland (thyroid dysgenesis), or a complete absence of thyroid tissue (athyreosis). The rest of 15% have thyroid hormonogenesis failure caused by inheritance. A daily iodine intake less than <25 μg, particularly in preterm infants, founds in many cases in Europe, Africa and Asia. Clinical diagnosis occurs in <5% of newborns have hypothyroidism because signs and symptoms are often minimal. As a result, without prompt diagnosis and treatment, most affected children progressively develop growth failure, irreversible mental retardation and a variety of neuro psychological deficits.

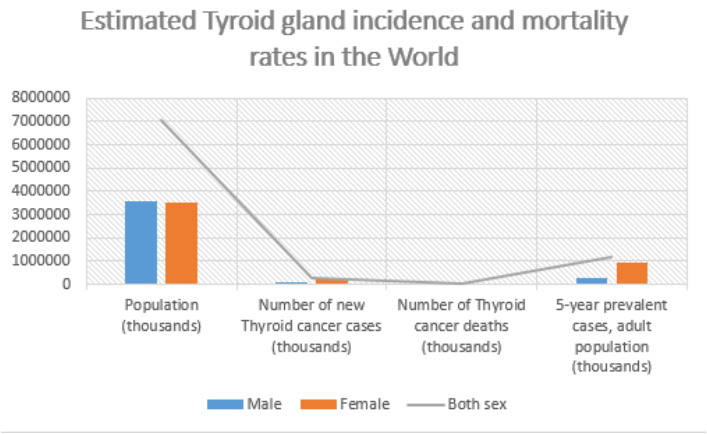


Figure 1: Estimated Thyroid gland incidence and mortality rate.

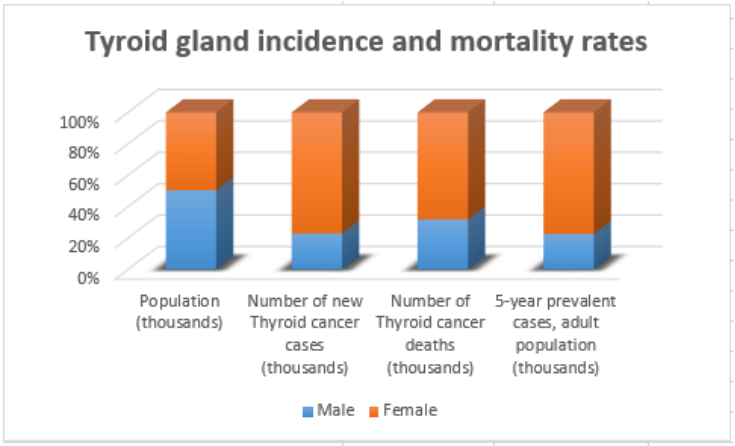


Figure 2: Estimated Thyroid gland incidence and mortality rate as percentage

Technology availability is also a challenge that stands for countries. Figure3 attempts to extrapolate the prevalence rate for Thyroid disorders to the populations of different countries and regions. These prevalence extrapolations of Thyroid disorders are only estimates, based on applying the prevalence rates from the United States to the population of any other countries. The prevalence of Thyroid disorders is 20 million Americans (National Women's Health Information Center (NWHIC), thus the prevalence Rate for

Thyroid disorders approx. 7.35% means1 in 13 or 20 million people in USA. Regarding to Undiagnosed Prevalence Rate for Thyroid disorders it is approx. 4.78% means 1 in 20 or 13 million people in USA.

As a result, lifetime risk for Thyroid disorders is 1 in 8 women during their lifetime in the United states according to (American Medical Women's Association); 1 in 8 for women (NWHIC) [11].

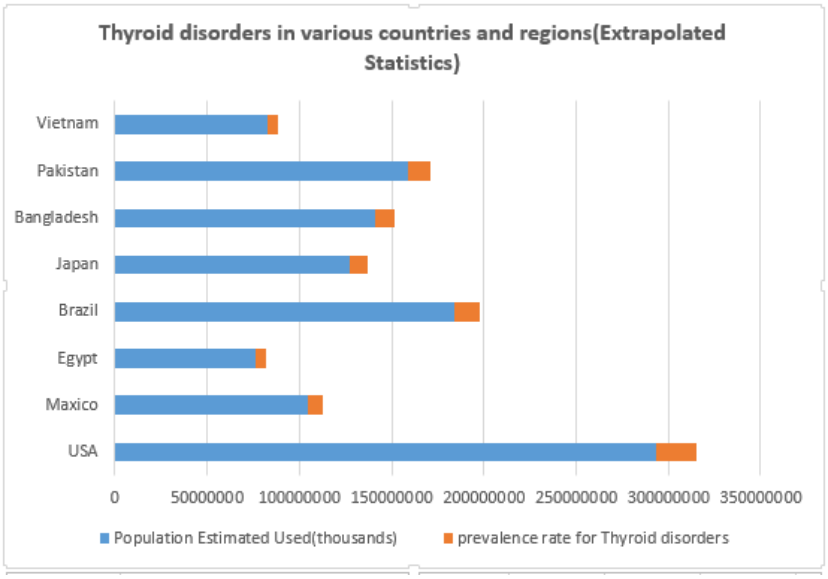


Figure 3: Prevalence rate for Thyroid disorders of different countries

While thyroid gland disease is one of the most commonly reported disease, for which data were available, worldwide, thyroid gland cancer was the eighth estimated age-standardized incidence and mortality rates in women.

For all the previous reasons, these shortages in medical resources and the serious effect of the disease especially on newborns motivates us to look for more effective solutions for the benefit of society. As well as computer engineers can utilize the latest technologies

in machine learning science to generate models and methods that can assist physicians in the process of diagnosis and treatment.

2.2 Research objectives

The aims of this research work are:

- To find a new way to utilize patient's histories, health information, and databases for detecting and diagnosing diseases, also provide predictive tool as medical professionals. This research is expected to establish an app that can assist physicians in diagnosing diseases and classifying patients in useful patterns based on different attributes, and how machine-learning techniques can be effective to identify such patterns. This can help in discovering early onset of the disease, treatment plans and identification of disease stages.
- To deal with large number of features and attributes in the dataset, and identify the significance of some features over others. However, large number of features can lead up to curse of dimensionality, also could render a machine learning algorithm or technique limited in terms of accuracy, specificity and precision.
- To address an important issue related to making up a GUI combining with machine learning algorithm, that can play an important role in determining the acceptability and ease of use achieved by designing technologies and machine learning algorithms.

Therefore, this thesis proposes new methods for investigate machine-learning techniques and develop machine-learning tool for providing predictive tool for thyroid gland disease diagnosis. This work envisages that the outcomes of this research in terms of an intelligent prediction app , with easy-use design of GUI , joining with decision tree algorithmic techniques, extracting significant 'feature selection' and using machine

learning based classification can reinforce the accuracy with which hyperthyroidism and hypothyroidism or a normal condition of the dysfunction can be identified. Figure 4 shows our research method overview.

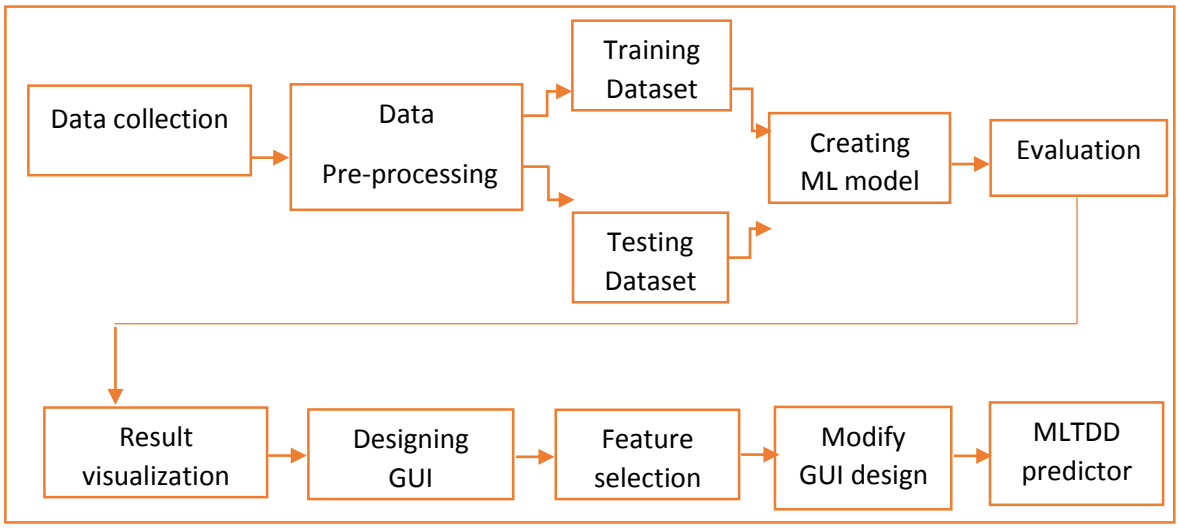


Figure 4: Research Method overview

To conclude, research objectives of this work is to utilize patient's histories, health information, and databases from collection of databases at repositories for discovering signs for early of thyroid gland diagnosis with an integrated intelligent prediction app consisting of significant feature selection, and learning based classification. The research is expected to establish a tool or an app that can assist physicians in diagnosing diseases. The aim is to design an ML app that combines the human expertise with the technology intelligence to achieve diagnosis that is more accurate. This app may assist physicians in decision making, for better and effective treatment plans.

2.3 Research Contribution

The aim of the current thesis is to analyze large data collected from university of California of Iravin (UCI) repository using machine-learning algorithms. The process of analyzing large amount of data includes some machine learning algorithmic techniques such as investigating better features selection techniques, and developing new machine learning based app for diagnosing the disease based on previous history information obtained from patients.

In terms of features selection techniques, the research applied features selection technique as a method to reduce the number of features in order to be able to design an acceptable GUI also to gain high quality attributes to improve the mining process. Features selection techniques effect all discipline that have need knowledge discovery from large data. The study found that features selection methods are capable of improving the interface of the proposed tool. However, features selection methods improve the performance of learning algorithms, and no single features selection method can best satisfy all datasets and learning algorithms. For this reason, machine-learning researchers should recognize the nature of datasets they studying and learning algorithm characteristics in order to obtain better outcomes.

In regards to diagnosis approaches, this work proposed an app for diagnosing thyroid gland based on machine intelligence and previous history. This app presented a method for thyroid gland diagnosis using a combination of a Classification and Regression Trees (CRT) and the feature selection method. In this approach, the purpose of CRT is to create a model that predicts the value of a target variable (class label) based on different input variables, and the purpose of information gain method is to reduce the number of input features. The experimental validation shows 99.7% accuracy, which confirms the capability of the proposed predictive tool.

With regard to programing the app, this work uses PyDev, which is a plugin that enables Eclipse to be utilized as a Python IDE ,supporting also Jython and IronPython [11].

It uses advanced type of inference techniques to provide features such code analysis and code completion, while still providing many others such as refactoring, a debugger, interactive console, etc. Combining with Qt Designer, which is Qt's tool for designing and building graphical user interfaces (GUIs) from Qt components. It enables compose and customize any widgets or dialogs in a what-you-see-is-what-you-get (WYSIWYG) manner, and experiment them using different styles and resolutions [12].

MLTDD is a useful tool to aid researchers in computer science and several other fields of science. Whether the target research is in medicine, business, agriculture, or industry; the requirement for analyzing large amount of data is needed. In addition to that, finding the most relevant feature selection technique that best satisfies a specific learning algorithm could benefits researchers. Therefore, the current work proposed a new app for diagnosis; the idea is to obtain an integrated app that combines the best performing learning algorithms technique and the best performing user Interface technologies with an experimental evaluation on the Ann thyroid gland Dataset. As a result using both human knowledge and machine learning ability offers the best performance.

Chapter 3

Background Study and Literature Review

3.1 Background study

Generally, machine learning can be defined as a scientific scope that aims to design and develop algorithms that allow computers to learn from data and manage to solve a real time problem based on previous stored data or under a particular instructions and rules. One of various presentations of machine learning is data mining, which is the most used application of machine learning [12]. Data mining is a science to discover knowledge from databases. The database contains a collection of instances (samples or records). Each sample used by machine learning algorithms is formatted using same set of fields (inputs or variables, features or attributes). The process of machine learning with the class label (the correct output) of instances then the learning process is called the supervised learning. In contrast, when the instances does not contain the class label then the learning process is called unsupervised learning process. Clustering is a common unsupervised learning method. The purpose of clustering is simply to describe data. However, our focus in the current research is on supervised machine learning classification and regression as predictive methods [12].

3.2 Machine learning

Machine learning (ML) can be defined as a group of topics that concentrate on making and testing algorithms that can improve the process of classification, prediction, and pattern recognition, by using computer models obtained from previously stored data (existing data). Machine learning can produce classifiers and predictors to be used on the available resources (repositories). In addition, machine learning does not involve much human interaction. The objective behind finite human involvement is that the use of

automatic pre-programmed methods can minify human biases. The goal of the process is proposing the algorithm and its functionality to classify objects or predict new cases, which based on tough and reliable data [13].

Machine learning can be defined as the science of making computers act without being actually programmed. In the past decade, machine learning has provided for us lots of effective apps such as practical speech recognition, effective web search and autonomous or self-driving cars, and a widely improved understanding of the human genome (complete set of nucleic acid sequence for humans).

In the next section, we will present some of the various applications of Machine learning and the most effective machine learning techniques. More importantly, the gain of the practical know-how needed to speedily and powerfully apply these techniques to solve new problems [14].

3.2.1 Snapshot Applications of Machine Learning

Machine learning systems are learning programs automatically from data. This is often a very interesting alternative to manually constructing them, in the last decade; the use of machine learning has spread speedily throughout computer science. Machine learning algorithms are being applied in various fields in interesting ways. It is becoming increasingly popular with more and more applications in areas where we may not even think of.
Below is a list of three of the most interesting applications.

- Protecting Animals

  An American research university in Ithaca, New York called Cornell University is working on an algorithm that can identify whales' location in the ocean based on

audio recordings in order to ships can avoid hitting them. Further, Oregon State University is developing a software that can define and determine which bird species is/are based on a given audio recording collected in field conditions.

- Predicting Emergency Room Wait Times

  Patients who require to be seen in 1 to 14 minutes are being seen in nearly twice that timespan in about 37 minutes, according to the (The U.S. Government Accountability Office GAO) in 2009. Even worse, the patients mast of the times get fed up of waiting and may leave without getting medical treatment. However, healthcare organizations and health tech companies are using a technique called Discrete Event Simulation that can predict wait times for patients at emergency department waiting rooms. The algorithm to create a model uses factors such as patient data, emergency department charts, staffing levels, and even the system of the emergency room to predict wait times.

- Identifying Heart Failure

  IBM with 1.28% researchers have found a way to read out heart failure diagnosis criteria from free-text physician notes. A machine learning algorithm is being developed in order to search through physicians free-form text notes in the (EHR) electronic health records and synthesize the text with "NLP Natural Language Processing". By a similar to the job of a cardiologist who can read through another physician's notes and determine whether a patient has heart failure, now computers can do the same task.

In the 2010s and 2020s, effectual, powerful analytics and machine learning are revolutionize industries, just as software revolutionized the world over the past 30 years [15].

3.3 Machine learning for Healthcare

Machine Learning is the most prevalent component of many creative software startups that are looking for re-define their markets. Similar to the way of healthcare, where great amounts of data are feeding medical scientists, drug makers and healthcare providers with information treasure that can be used to derive observation. In this part, we have selected four from several highly ranked healthcare companies that are innovating using machine learning.

MedAware offers innovative resolutions that save lives, significantly decrease healthcare costs and improve patient safety; the idea is based on identifying and preventing errors on prescription in real-time. MedAware's technology uses machine learning algorithms with big data analytics to analyze large targeted data of Electronic Medical Records (EMRs), to learn automatically how doctors treat patients in real life.

Ginger.io is based on a predictive models; the focus of this company is on people with depression and anxiety. It uses a combination of data science, smartphone technology, and clinical services to create a personalized, fair way to deliver mental healthcare. With this company, patients will be able to psychoanalyze their mood over time, and receive extra mental health support as needed.

Wellframe is a startup company focused on building the following generation of infrastructure for healthcare delivery system using machine learning that offers a mobile-enabled care management platform. The company's partners with leading health

providers and healthcare stakeholders to develop care management services to assist patients forward improved experience, care plan commitment and health outcomes.

Lumiata provides predictive health analytics; it has developed the world's first medical graph. This medical graph analyzes and organizes hundreds of millions of valuable data points, helping Lumiata to deliver real-time predictive analytics that assist hospital networks and insurance carriers to present higher quality care to extra patients in less time. It provides accurate predictions related to procedures, symptoms, medications, and diagnoses.

As machine learning techniques are applied to massive data that has appeared in recent years, there is the potential for massive innovations in healthcare. We will see how these four current leaders, and other startups, progress in the future [16].

3.4 Machine Learning Algorithms

In this part of research, the most popular machine learning algorithms has been presented. It is important to display the main algorithms in order to know the methods available. It can feel confused when algorithm names are available, but this part is expected to just give a brief description of what they are and where they fit.

There are two approaches to categorize the algorithms; the first approach is a grouping of algorithms by **similarity** in function or form. The second approach is a grouping of algorithms by the **learning style**. Both methods will be presented as well as variety of different algorithm types.

3.4.1 Algorithms Grouped by Learning Style

Different ways are available for an algorithm which can follow to model a problem; based on its input data which can be called as environment or interaction with the experience.

There are three main learning styles (learning models) and they will be presented in this section with some examples of algorithms. This way of organizing machine-learning algorithms (taxonomy) is useful because it forces us to think about two important factors which are; the roles of the input data and the model making process and then select the most suitable for the problem to get the best result. The following are three different learning styles in machine learning algorithms:

First: Supervised Learning

In supervised learning the input data or training data has a known target variable. A model is prepared through a process called (training process) then it is used to make predictions, as shown in figure 5. The training process goes on until a good accuracy can be obtained.

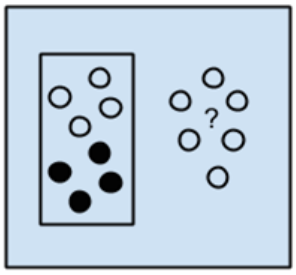


Figure 5: Supervised Learning algorithms

The main two problems are classification and regression. Example algorithms include Support Vector Machines , Nearest Neighbors, Logistic Regression and Naive Bayes .Supervised Learning and classification problem will be presented in details in the next section.

Second: Unsupervised Learning

In unsupervised learning, the Input data does not have class label or a known result, as shown in figure 6.

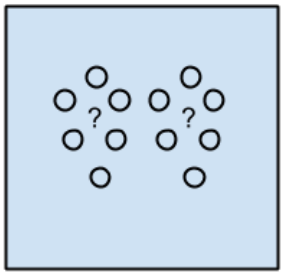


Figure 6: Unsupervised Learning algorithms

Third: Semi-Supervised Learning

In semi-supervised learning, some of the input data has a target variable and the rest is unlabeled samples. The model has to learn the structures to organize the data in order to make predictions. It deals with classification and regression problems.

3.4.2 Algorithms Grouped By Similarity

Algorithms are grouped by similarity according to their function (how they work). This is the most useful way (approach) to group algorithms. In this section, we present several of the common machine leaning algorithms grouped by similarity in eleven different groups:

- Regression Algorithms
- Instance-based Algorithms
- Regularization Algorithms

- Decision Tree Algorithms
- Bayesian Algorithms
- Clustering Algorithms
- Association Rule Learning Algorithms
- Artificial Neural Network Algorithms
- Deep Learning Algorithms
- Dimensionality Reduction Algorithms
- Ensemble Algorithms [17].

3.5 Classification

Classification and regression are the most popular models in supervised learning. This research will concentrate on classification. However, it is necessary to distinguish between them. While regression algorithms attempt to predict the target values of the instances, most probably would be continues value. Regression is used to model the relationship between a target variable and a set of predictor variable. This relationship is considered as a function that predicts the target variable using the predictor variables as inputs. On the other hands, classifiers can categories the input instances into pre-defined classes, for instance, a classifier can predict a new patient if it is healthy or malignant (suffer from a certain disease) [18].

Classification is the process of learning the goal function that maps between a set of instances (inputs) and a predefined class labels (output target). The input data for the classification is a set of samples, which called instances. Each instance is a record of data in the form of ($\boldsymbol{x}$, y) ,where $\boldsymbol{x}$ is the features set and $\boldsymbol{y}$ is the target variable (class label).

Classification model can be a tool that used to describe data (Classifier or Descriptive Model) or a tool to predict the class or target variable for new instance (Predictive Model) [19].

The overall approach for solving classification problem is shown in Figure 7. The training data contains instances whose output or class labels are known. The classification model can be built based on the training data. The model then can be tested and evaluated by using the testing data which consists of records with missing class labels. The evaluation of model performance is based on the number of testing instances that are correctly predicted [19]. The outcome of performing the model on the testing data produces the confusion matrix.

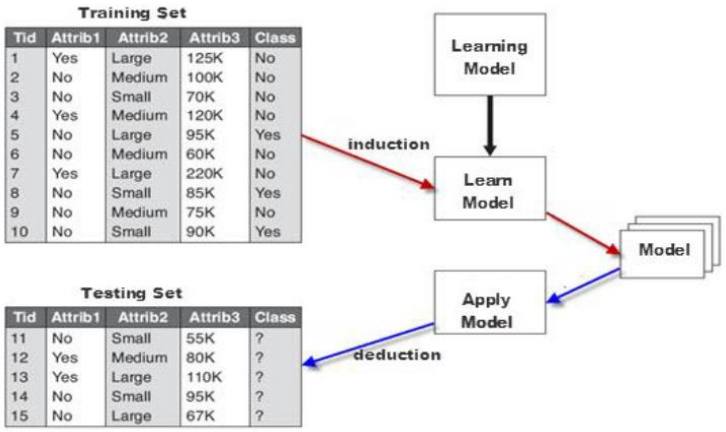


Figure 7: General approach for building a classification model

Assume the goal is to classify some objects $\boldsymbol{i}$=**1**,…., $\boldsymbol{n}$ into $\boldsymbol{k}$ predefined classes, where $\boldsymbol{k}$ represent the number of classes. For example, if the aim of classification is to diagnose a patient whether or not suffering from thyroid gland disorder then the value of $\boldsymbol{k}$ will be 2 corresponding to either thyroid disorder or normal.

Database (available data) can be organized as $\boldsymbol{n\ x\ p}$ matrix $\boldsymbol{X}$, where $\boldsymbol{xij}$ represent the feature value $\boldsymbol{j}$ in the record $\boldsymbol{i}$. Every row in the matrix $\boldsymbol{X}$ is represented by a vector $\boldsymbol{xi}$ with $\boldsymbol{p}$ features and a class label $\boldsymbol{yi}$. The classifier can be denoted as ($\boldsymbol{x}$). One method to evaluate the classifier is by calculating the error estimation based on the confusion matrix. To clarify and explain the error estimation, let us consider an example. Suppose the aim of a certain classifier ($\boldsymbol{x}$) is to train and test input vectors $\boldsymbol{x}$ into two possible classes normal and disorder. Suppose the result of classification of the classifier ($\boldsymbol{x}$) on vectors $\boldsymbol{x}$ is as shown in the confusion matrix in Table 1.

	Predicted	
True	normal	disorder
normal	60	14
disorder	6	80

Table 1: The confusion matrix for classifier cl(x) on matrix X

Er is the error rate of algorithm, which calculated as the total number of incorrect classified samples divided by the total number of records in the matrix X. In our example, Er = (14 + 6) / 160 = 0.125. At the same time, to calculate the classification accuracy of the model

Accur = 1 − Er = 0.875 can be applied.

Chapter 4

Research methodology

4.1 Thyroid gland disorder

This part of research presents, thyroid disorders, which are conditions that affect the thyroid gland. A description of Thyroid gland and how does it work has been explained as well as the thyroid hormones effectiveness and how the thyroid has significant roles to regulate many metabolic processes throughout the body. Various types of thyroid disorders affect either its function or structure.

4.1.1 Thyroid gland system

The thyroid gland is an organ located in the base of your neck; it is a butterfly-shaped. It releases hormones that control metabolism: Metabolism—the way your body uses energy— this involves heart rate and how quickly the body uses calories from the eaten food . Metabolism can be interpreted as a collection of chemical reactions that takes place in the body's cells. The process of metabolism converts the fuel in the food into the energy needed to power our activities, from growing to moving to thinking. The chemical reactions of metabolism controlled by specific proteins in the body, and each chemical reaction is coordinated with a particular body functions. In fact, thousands of metabolic reactions happen to keep our cells working and healthy. The thyroid's hormones regulate numerous body functions, including:

- Breathing
- Muscle strength
- Heart rate

- Cholesterol levels
- Body weight
- Body temperature
- Menstrual cycles

The thyroid gland longs about 2-inches and lies in front of your throat. The thyroid has two sides known as lobes that lie on either side of your windpipe, and is connected by a strip of thyroid tissue called an isthmus. Some people have two separate thyroid lobes instead of an isthmus [20].

The endocrine system, which is made up of glands that produce, store, and release significant hormones into the bloodstream in order to enable hormones to reach the body's cells. The thyroid is part of the endocrine system and it gets iodine from the foods we eat and uses it to make two main hormones:

- Triiodothyronine (T3)
- Thyroxine (T4)

It is so important that T3 and T4 levels are in regular level (neither too high nor too low). Two glands in the brain—the hypothalamus and the anterior pituitary gland communicate to maintain T3 and T4 balance.

The hypothalamus produces TSH Releasing Hormone (TRH) that sends signals the pituitary to stimulate the thyroid gland to produce more or less amount of T3 and T4 by increasing or decreasing the release of (TSH) thyroid stimulating hormone as shown in Figure 8.

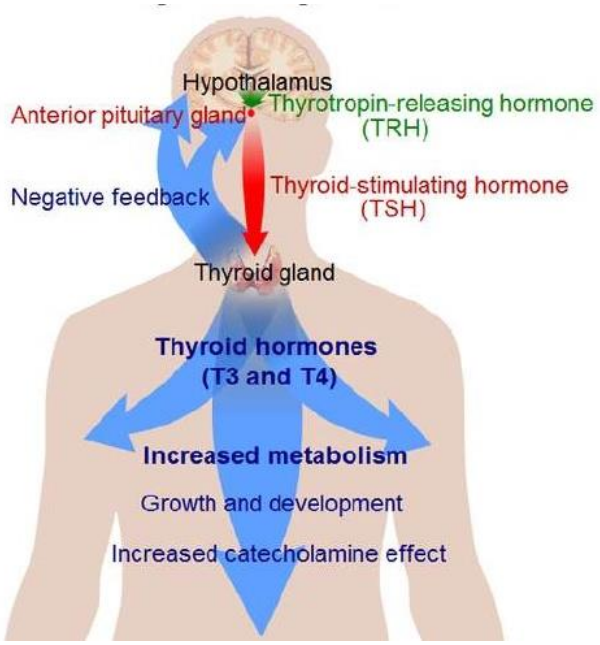


Figure 8: Thyroid system

- If T3 and T4 levels are low in the blood, the pituitary gland releases more TSH to stimulate the thyroid gland to produce more thyroid hormones.
- When T3 and T4 levels are high, the pituitary gland releases less TSH to stimulate the thyroid gland to slow production of these hormones.

4.1.2 Effects of Thyroid Hormones

T3 and T4 travel in bloodstream to reach almost every cell in the body. The hormones regulate the speed with which the cells or the metabolism works. For example, T3 and T4 regulate heart rate. Therefore, if T3 and T4 levels are low, the heart rate may be slower than normal; also that low level may cause constipation/weight gain. If T3 and T4 levels are high, a person may have a rapid heart rate and diarrhea/weight loss.

4.1.3 Thyroid Diseases

There are specific kinds of thyroid diseases that includes:

- **Functional disorders**

  - Hypofunction - Hypothyroidism

    Hypothyroidism, also well known as underactive thyroid disease, is a common disorder in which thyroid gland does not produce enough amount of thyroid hormone. It can cause a number of symptoms, such as weight gain, a feeling of tiredness, depression, and poor ability to tolerate cold. Occasionally there may be swelling in the front side of the neck because of goiter. Untreated hypothyroidism during pregnancy can cause delays in growth and intellectual development in the baby, which is known as cretinism.

    Worldwide, most common cause of hypothyroidism is too little iodine in the diet. In countries where enough iodine in the diet, the most common cause of hypothyroidism is (Hashimoto's thyroiditis) the autoimmune condition. Some less common causes are: injury to the hypothalamus or the anterior pituitary gland, previous treatment with radioactive iodine, certain medications, previous thyroid surgery or a lack of a functioning thyroid at birth. The diagnosis of hypothyroidism can be approved with blood tests measuring (TSH) thyroid-stimulating hormone and thyroxine levels.

  - Hyperfunction - Hyperthyroidism

    Hyperthyroidism is a condition of the thyroid produces too much of one of the two primary hormones (tetraiodothyronine T4 and triiodothyronine T3) or both. This can cause elevated blood pressure, rapid heart rate and hand tremors, as well as many other symptoms. Diagnosis of the overactive thyroid and treatment of the underlying cause can minify symptoms and stop complications. Several conditions can cause hyperthyroidism. The most

common cause is Graves' disease, which is an autoimmune disorder where antibodies stimulate the thyroid to secrete too much hormone. It occurs more often in women than in men. Other causes of hyperthyroidism can include: excessive iodine (the main key ingredient in T4 and T3) / inflammation of the thyroid (thyroiditis) that lead to leaking of T4 and T3 of the gland / benign tumors of the thyroid or pituitary gland / large amounts of tetraiodothyronine taken from medication or dietary supplements.

- **Nodular abnormalities - Goiter**

A goiter is an enlarged thyroid gland. When the gland turns into enlarged because of diseases or tumors, then it is referred to as a goiter. That can be associated with over-function of the thyroid gland or with under-function of the gland. In addition, both inflammation and tumors can cause a goiter. Sometimes, the whole gland may be enlarged, while in other goiters, nodules, or enlargement may occur in one part of the gland only.

- **Thyroid nodules**

Nodules are lumps that generally arise within a normal thyroid gland. Commonly these abnormal growths can be felt as a lump in the throat since are located at the edge of the thyroid gland.

- **Deficiencies**

Hashimoto's disease: Is an autoimmune disease, a disorder in which the immune system turns against the body's tissues. In patients with Hashimoto's, the immune system attacks the thyroid gland and this can lead to several deficiencies on the gland.

- **Thyroid cancer**

Cancer of the thyroid gland is a disease in which malignant cells are found in the tissues of the thyroid gland. Thyroid cancer is more common in women than in men. Age average of most patients is between 25 and 65 years old. Patients who have had

radiation treatment due to medical problems in the neck or head so they have been exposed to large amounts of radiation have a higher chance of getting cancer on thyroid gland [21].

4.2 Description of data set

4.2.1 General Description

This research relied an option of utilizing online databases. Online databases are collected from clinical environment and are available publicly, freely for research proposes. Furthermore, they have undergone proper organizational ethics approval processes. The advantage of using online databases is the ability to compare our work and the existing works by using the same databases.

UCI machine learning repository is one of the most popular machine learning repositories, which is a collection of data generators and databases that are used by machine learning researchers to train and test algorithms of machine learning. The repository was created in 1987 by David Aha and fellow graduate students at University of California, Irvine (UC Irvine). Since that time, this repository has been widely used by educators, students, and researchers all over the world as a reliable source of machine learning databases.

This research has used the dataset (ann-thyroid dataset) from UCI repository. It consists of 7200 instances and (3) classes, 3772 are training instances, 3428 testing are instances and 21 attributes as shown in Table 2 [23]. The task is to detect whether a given patient has a normal condition (1) or suffers from hyperthyroidism (2) or hypothyroidism (3).

Thyroid Disease (ann) data set			
Type	Classification	Origin	Real world
Features	21	(continuous / categorical)	(6 / 15)
Instances	7200	Classes	3
Missing values			No

Table 2: Thyroid Disease (ann) data set General information

Figure 9 shows a code in python language has been used initially build in at our project to account the number of patients who have normal condition of thyroid functionality and patients with hyperthyroidism as well as patient with hypothyroidism.

```
print "-----------------------------------"
print 'The number of instances per class'

n = 0
hr = 0
ho = 0

for k in range(0, 7200):
 if y[k] == 1 :
      n = n + 1
 elif y[k] == 2 :
      hr = hr + 1
 elif y[k] == 3 :
      ho = ho + 1


print 'class 1 : normal = ', n
print 'class 2 : hyper = ', hr
print 'class 3 : hypo = ', ho
print "-----------------------------------------
```

```
-----------------------------------
The number of instances per class
class 1 : normal =  166
class 2 : hyper =  368
class 3 : hypo =  6666
-----------------------------------
```

Figure 9: Python code to get the number of instances per class

4.2.2 Attribute Description

This section describes the main characteristics of the thyroid data set and its attributes: Each measurement vector consists of 21 values – 15 binary and 6 are continuous. The binary attribute values are mapped to zero and one where zero refers to False and one refers to True. Each of the measurement vectors belongs to one of three different classes, which are either hyper-thyroidism, hypo- thyroidism or normal function of the thyroid gland (Table 3).

Attribute	Domain
1. Age	[0.01, 0.97]
2. Sex	[0, 1]
3. On_thyroxine	[0, 1]
4. Query_on_thyroxine	[0, 1]
5. On_antithyroid_medication	[0, 1]
6. Sick	[0, 1]
7. Pregnant	[0, 1]
8. Thyroid surgery	[0, 1]
9. I131_treatment	[0, 1]
10. Query_hypothyroid	[0, 1]
11. Query_hyperthyroid	[0, 1]

12. Lithium	[0, 1]
13. Goiter	[0, 1]
14. Tumor	[0, 1]
15. Hypopituitary	[0, 1]
16. Psych	[0, 1]
17. TSH	[0.0, 0.53]
18. T3	[0.0005, 0.18]
19. TT4	[0.002, 0.6]
20. T4U	[0.017, 0.233]
21. FTI	[0.002, 0.642]
Class	**{1,2,3}**

Table 3: Thyroid Disease (ann-thyroid) data set Attribute description.

The table above illustrates the domain of each attribute, python language has been used to find out the minimum and maximum value of each attribute, figure 10 shows that.

```
print 'The domain of the Age feature is==>'
b = m[:,0]
print min(b)
print max(b)


for i in range(16, 21):
    b = m[:,i]
    print '--------------------------------'
    print 'The domain of the feature ', i+1,'is==>'
    print min(b)
    print max(b)
    print '--------------------------------'
```

Figure 10: Python code to get the domain of each attribute

4.2.3 Data set visualization

The following figure (Figure 11) shows visualization for the dataset, it has been observed that from 7200 instances there are 166 normal, 368 hyperthyroidism and 6666 hypothyroidism.

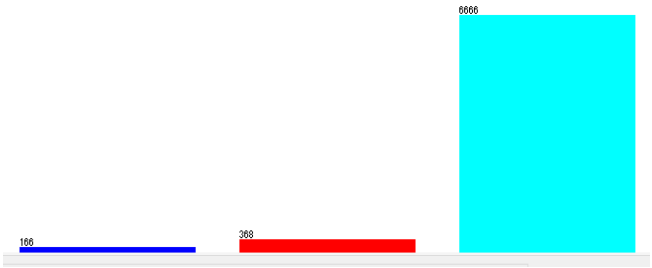


Figure 11: Class visualization for the whole dataset

As a result, in this particular dataset most patients have a hypothyroidism disease whereas a small number of them have a normal thyroid gland condition.

4.3 Decision Tree algorithm

Always a code is needed to tell a computer what to do but an algorithm is necessary before writing that code. An algorithm can be defined as a list of rules to follow in order to solve a problem.

4.3.1 Decision tree representation

A decision tree is a flowchart it has a tree structure, where each interior node (non-leaf node) represents a test on attribute, every branch symbolizes an outcome of the test, the nodes that holds a class label called leaf node or (terminal). The topmost node known as (the root node). The goal is to create a model that predicts the value of a target variable based on input variables.

A tree is "learned" by splitting the origin data set into subsets based on an attribute value test. Then this process is done on each obtained subset in a recursive manner know n as (recursive partitioning). The recursion ended whenever the subset at a node has the same value of the class variable. This kind of process is called (TDIDT top-down induction of decision trees) , it is the most common approach for learning decision trees from data.

Decision trees can be described on data mining as the combination of mathematical and computational techniques to extract the description, categorization and generalization of a given set of data.

The following form is how data comes in records:

$$(\mathbf{x}, Y) = (x_1, x_2, x_3, ..., x_k, Y)$$

The dependent variable, Y, is the class label (target variable) that we are trying to understand, generalize and classify. The vector x is consist of the input variables, x1, x2, x3 etc.., which are used for that task [25].

4.3.2 Decision tree types

There are two main types of decision trees:

- **Classification tree**: the goal is to predict the class to which the data belongs.
- **Regression tree**: the predicted outcome considered as a real number.

The term Classification And Regression Tree (CART) is an overall term used to refer to both of the above procedures. Regression trees and classification trees have some similarities and some differences as well.

Some techniques, known as ensemble methods, structure and builds multiple decision trees (more than one) such as Bagging, A Random Forest, Boosted Trees and Rotation forest .

Decision tree learning is the construction of a decision tree from class-labeled training tuples. Decision list is a special case of a decision tree, which is a one-sided decision tree where each internal node has exactly 1 leaf node and only 1 internal node as a child unless the bottommost node, where a single leaf node is its only child . Decision lists are easier to understand than general decision trees .

Here are some notable decision-tree algorithms:

- **ID3** (Iterative Dichotomiser 3)
- **C4.5** (successor of ID3)

- **CART** (Classification And Regression Tree)

CART and ID3 were invented at nearly the same time (between 1970 and 1980), and follow a similar approach which is learning decision tree from training data [25].

4.3.3 Decision tree Metrics

For constructing decision trees, the algorithms usually work top-down, by selecting a variable at each step that best splits the set of data (items). Various algorithms use different metrics in order to measure "best". These generally focus on measuring the homogeneity of the target variable with the subsets. By applying these metrics to each candidate subset, the resulting values are combined to provide a measurement of the quality of the split.

The first metrics is Gini impurity

Generally used by the CART (classification and regression tree) algorithm, it is a measure of how often a (randomly chosen element) from the set would be incorrectly labeled. Gini impurity can be computed by summing the probability f_i of each item being chosen times the probability $1 - f_i$ of a error in categorizing that item.

We can see from Eq. (1) that to compute Gini impurity in a set of m items, assume $i \in \{1, 2, ..., m\}$, and let f_i be the portion of items labeled with value i in the set.

$$I_G(f) = \sum_{i=1}^{m} f_i(1 - f_i) = \sum_{i=1}^{m}(f_i - {f_i}^2) = \sum_{i=1}^{m} f_i - \sum_{i=1}^{m} {f_i}^2 = 1 - \sum_{i=1}^{m} {f_i}^2 = \sum_{i \neq k} f_i f_k \quad (1)$$

The second metrics is Information gain

Commonly used by the ID3, C4.5 algorithms. It is based on entropy concept from information theory.

Defined as in Eq. (2):

$$I_E(f) = -\sum_{i=1}^{m} f_i \log_2 f_i \tag{2}$$

From Eq. (3) we calculate Information Gain = Entropy (parent) - Weighted Sum of Entropy (Children)

$$IG(T, a) = H(T) - H(T|a) \tag{3}$$

4.3.4 Popular Decision Tree Classification & Regression Trees (CART)

In this part of research, we discuss the CART decision tree (Classification and Regression Tree), classification and regression trees are common machine-learning methods with the objective of constructing (Prediction models) that predicts the value of a dependent variable, which called target according to the values of independent variables (input data).

In 1984 the CART Trees methodology was introduced, and that was by Leo Breiman, Jerome Friedman, Richard Olshen and Charles Stone as an overall term to refer to the two types of decision trees:

- Classification Trees: Generally, the target variable is categorical (sometimes called a nominal variable) is one that has two or more categories, and the tree is used to identify the "class" for which a target variable belongs to.
- Regression Trees: The target variable is mostly continuous which means that the variable can use any value between its minimum and its maximum, and tree is used to predict its value.

The structure of CART algorithm is like a sequence of questions, so that the answers to which determine what the next question. The outcome of these questions is a structure

shapes like tree where the ends called terminal nodes at any point where no more questions [27].

Here are some useful features and advantages of CART:

- CART has the capability, which enable it to use the selfsame variables more than one time in different parts of the tree.
- CART is nonparametric so it does not rely on data belonging to a specific type of distribution.
- CART combines both testing with a test data set and cross-validation to estimate the goodness of fit more accurately.
- CART can be used with other prediction methods to pick the input set of variables.
- Outliers in the input variables do not significantly affect CART [28].

CART classification algorithm which is based on decision tree induction (Jiwai H. and Micheline Kamber,2009) which is a learning of decision trees from class label training tuples. The Classification and Regression (CART) tree method uses recursive partitioning to split the training records data into segments with analogous output field values. The CART tree node begins by examining the input fields to detect the best split, scaled by the reduction in an impurity index that outcomes from the split. CART uses Gini index splitting records measures in choosing the splitting attribute. Then pruning is done in CART by using a training data set. That split defines two subgroups; each of them will subsequently split into two more subgroups, until one of the stopping standard is moved. All splits are binary (only two subgroups).

Algorithm 1 has been used for the construction of the decision tree [29].

- Input is Data partition say (D) which is (training dataset).
- Attribute list is Attribute selection method.
- Output will be Decision tree.

- Create a root node assume(N);
- If tuples in (D) are all of the same class, assume (C) then
  Return (N) as a leaf node labeled with the class (C);
- If the attribute list is empty, then
  Return (N) as a leaf node labeled with the majority class in (D);
- Apply feature selection method to find the best splitting rule;
- Label node (N) with splitting criterion;
  Attribute list equal to attribute list – splitting attribute;
- Then For each outcome assume 'j' of splitting criterion
  Let 'Dj' be the set of data tuples in (D) satisfying outcome j;
- If 'Dj' is empty, then
  Attach a leaf node labeled with the majority class in (D) to root node (N);
  Else, attach the node returned by producing decision tree to root node (N);

End for
Return (N) as a root node;

Algorithm 1: Algorithm used for the construction of the decision tree

Chapter 5

Developing a ML tool for Thyroid Disease Diagnosis (MLTDD)

5.1 Machine learning in Python

Machine Learning in Python displays how to achieve a successful data analyzing using machine learning algorithms, also how to apply them using Python [30].

5.1.1 Python programming language

Python is an open source programming language; it has two important advantages, look good and easy to read. Guido van Rossum is the programmer who made it in 1991. It is named after "Monty Python's Flying Circus" which is a television show.

Python is an interpreted language. The term (Interpreted languages) refers to kind of languages which do not need to be compiled to run. An interpreter, which is a special program, will run python code on any kind of computer. Another feature; if the programmer needs to change the code they can immediately see the results. On the other hand, Python is slower than a compiled language like C or C++ because it is not running machine code directly. However, Python is a good programming language for beginners. Writing programs in Python takes less time than in another language. It is a high-level language, where a programmer can focus on what to do instead of how to do it. Python got inspiration from previous programming languages like C, C++, Java, Perl, and Lisp. Python 2.7 has been used for this project [31].

5.1.2 Python IDE (Integrated development environment)

Python programmer mainly looking for a robust library that can be used to bring machine learning into a production system, in this work the library that has been seriously considered is spyder.

One of the Scientific Python Development Environment for python is spyder; with advanced editing, reactive testing, debugging and examination features, it is definitely a powerful interactive development environment for the Python language. Furthermore, it is a numerical computing environment because of the support of IPython and popular Python libraries such as NumPy, SciPy or matplotlib.

There are various libraries working with spyder, here is a list of the most common libraries:

- NumPy (Numerical Python): Base n-dimensional array package, NumPy is the essential package for scientific computing with Python.
- SciPy (Scientific Python): Fundamental library for scientific computing.
- Matplotlib: Matplotlib is a Comprehensive and interactive python 2D plotting library, which makes figures in interactive environments through platforms. Matplotlib can be used in web application servers, python scripts, the python and IPython, and six graphical user interface toolkits.
- IPython IP[y] (interactive Python interpreter): Enhanced interactive console which provides a rich architecture for interactive computing with:
    - A powerful interactive shell.
    - Support for interactive data visualization and use of GUI toolkits.
    - High performance and easy to use tools for parallel computing.
- Pandas: *Pandas* is an open source, data structures and analysis library adding easy-to-use, high-performance, data analysis and data structures tools for the Python programming language.
- Scikit-learn: is a free software machine-learning library for the Python programming language. It presents various classification, regression and clustering algorithms involving support vector machines, gradient boosting and

k-means, in addition to its design in order to interoperate with the Python scientific and numerical libraries SciPy and NumPy.
Scikit-learn in general written in Python, with some core algorithms (algorithms that are designed to process too large amount of dada and allow it to fit into a computer's main memory at one time) written in Cython to achieve performance.

Spyder 2.3.7 has been used in this work, pandas, Scikit-learn, NumPy and matplotlib has been used in this project [32].

5.2 Applying CART algorithm

At this stage, the data are ready for applying the algorithm. The testing data will be used to evaluate the classifier. This work proposed CART algorithm.

5.2.1 Algorithm fitting

The first step is to import all the required Python libraries to the program (spyder) as shown in figure 12.

```
import pandas as pd
import numpy as np

import sklearn
import sklearn.datasets
import sklearn.cross_validation
from sklearn.cross_validation import train_test_split
```

Figure 12: importing the required Python libraries

The next step is to load the thyroid data set and store the input data and target values in X and Y variables. The X value has 7200 rows and 21 columns, and Y value has 7200 rows and 1 column.

Figure 13 shows how to convert thyroid.data into a pandas data frame. At first print the column names are just numbers, as it can be seen replacement those numbers with the feature names has been done.

```
thyroid = pd.io.parsers.read_csv(
    'D:/datas/ann-all.csv',
    header=None
    )

thyroid.columns=['Age', 'Sex', 'On_thyroxine', 'Q_on_thyroxine',
                 'On_antithyroid', 'Sick', 'Pregnant', 'Thyroid_surgery',
                 'I131_treatment', 'Query_hypothyroid', 'Query_hyperthyroid',
                 'Lithium', 'Goitre', 'Tumor', 'Hypopituitary', 'Psych', 'TSH',
                 'T3', 'TT4', 'T4U', 'FTI', 'Class']
print thyroid.head()
```

Figure 13: Converting the dataset into a pandas frame

The next step is to print out the dataset as shown in figure 14.

```
    Age  Sex  On_thyroxine  Q_on_thyroxine  On_antithyroid  Sick  Pregnant  \
0  0.73    0             1               0               0     0         0
1  0.24    0             0               0               0     0         0
2  0.47    0             0               0               0     0         0
3  0.64    1             0               0               0     0         0
4  0.23    0             0               0               0     0         0

   Thyroid_surgery  I131_treatment  Query_hypothyroid  ...    Goitre  Tumor  \
0                0               1                  0  ...         0      0
1                0               0                  0  ...         0      0
2                0               0                  0  ...         0      0
3                0               0                  0  ...         0      0
4                0               0                  0  ...         0      0

   Hypopituitary  Psych      TSH     T3    TT4    T4U    FTI  Class
0              0      0  0.00060  0.015  0.120  0.082  0.146      3
1              0      0  0.00025  0.030  0.143  0.133  0.108      3
2              0      0  0.00190  0.024  0.102  0.131  0.078      3
3              0      0  0.00090  0.017  0.077  0.090  0.085      3
4              0      0  0.00025  0.026  0.139  0.090  0.153      3

[5 rows x 22 columns]
---------------------------------------------------------------------
```

Figure 14: Sample of Ann-Thyroid Diagnosis dataset

Matplotlib library has been used to visualize class label and create a histogram of ann dataset as sown in figure 15.

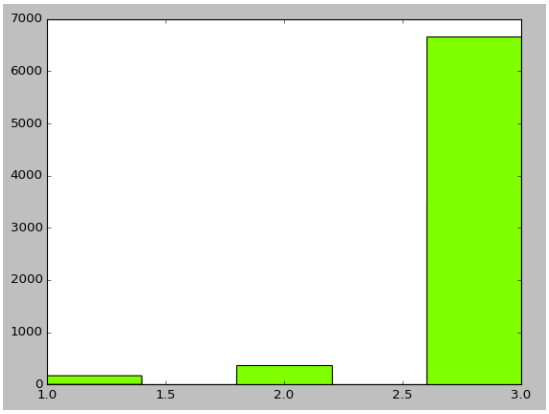


Figure 15: Histogram for Ann_tyroid dataset

Usually a dataset is divided into two datasets training and testing (sometimes it is called a validation set ' instead), a training set is implemented to train the classifier (to build up a model), whereas a test (or validation) set is to validate the model built (to test the performance of your classifier). Applying heuristics such as 10% testing and 90% training can bias the classification results and the results may not be generalizable. However, the training set can be nominated by applying a random filter to the data, for instance, select 80% of the points at random to generate the model and test against the remaining 20%. An accepted method is N-Fold cross validation, in which the dataset is been randomized and create N (nearly) equal size partitions. Then select Nth partition as testing and N-1 partitions for training the classifier. The perfect method is to employ another K-fold cross validation to generate a validation set and obtain the best parameters and repeat this process N times to find out an average of the metric. To be especially careful, doing this multiple times is required, by selecting different random training sets and compare the models obtained.

Sklearn cross validation is a function Split arrays or matrices into random train and test subsets, Sklearn cross validation has been implemented as a method of splitting Thyroid dataset into training and testing datasets as follows:

```
from sklearn.cross_validation import train_test_split

X_train, X_test, y_train, y_test =
sklearn.cross_validation.train_test_split(X, thy.CLAA, test_size=0.23,
random_state=12345)
```

The first parameter is test_size can be float, int, or none the default is None. If it is float, then the value will be between 0.0 and 1.0, this parameter represents the proportion of the test split of the dataset. Else if int, then it represents the absolute number of test samples. If it is none, then the complement of the train size is automatically set to the value. The last option is when train size is also none; then test size is set to 0.25. The second parameter is random_state can be int or RandomState which is Pseudorandom number generator used for selecting random sampling. In this work, 23 % for testing and 77 % for training has been selected. Therefore, 5544 Training samples and 1656 testing samples has been obtained.

Importing decision tree from Scikit learn module is required after dropping the last column as the parameters X values the only needed. then by storing decision tree object in a variable called (dt) fitting the CART algorithm is done therefore the classifier model is created.

```
from sklearn.tree import DecisionTreeClassifier
dt = DecisionTreeClassifier(criterion='entropy', splitter='best',
max_depth=20, min_samples_split=20, min_samples_leaf=1,
min_weight_fraction_leaf=0.0, random_state=None)

dt.fit(X_train, y_train)
```

Multiclass or multinomial classification has been our problem in the current work , In machine learning it is the problem of classifying instances into one of the more than two classes(also well-known as binary classification). Multiclass classification is completely different from multi-label classification, in multiclass each training point belongs to one of N different classes. The aim is to construct a function which, by given a new data point, can correctly predict the class label to which the new data belongs. Whereas, multiple labels means multiple class labels must be assigned to each new instance.

5.2.2 Evaluation

The goal is Learning and Predicting which means to make the model learn from our dataset's data and then to predict on new data. In scikit-learn, we learn from actual data by generating an estimator as follow:

Pred_train = dt.predict(X_train) Pred_test = dt.predict(X_test)

After predicting the class label for x values of training dataset and for testing dataset ,a comparison between the predicted values and the available values has been done in order calculate the classification and prediction accuracy as shown in figures 16,17.

```
print('\n Classification score for the training dataset')
print dt.score(X_train, y_train)

print('\n Classification score for the testing dataset')
print dt.score(X_test, y_test)
print "-----------------------"

pred_train = dt.predict(X_train)
print('\nPrediction accuracy for the training dataset')
print('{:.2%}'.format(metrics.accuracy_score(y_train, pred_train)))

pred_test = dt.predict(X_test)
print('\nPrediction accuracy for the test dataset')
print('{:.2%}\n'.format(metrics.accuracy_score(y_test, pred_test)))
```

Figure 16: Predicting the class label for training and testing dataset

```
 Classification score for the training dataset
0.996212121212

 Classification score for the testing dataset
0.998188405797
-----------------------

Prediction accuracy for the training dataset
99.62%

Prediction accuracy for the test dataset
99.82%
```

Figure 17: Classification and prediction accuracy

In regards to evaluation classifier based on classification report, precision, Recall, F1 score and support has been studied. In a classification task, the "Precision" for a class is the number of the true positives divided by the summation of the true positives and false positives. Whereas "Recall" is acquired by dividing the number of true positives by the full number of elements that actually belong to the positive class, in other words the sum of true positives and the items that not belonging to the positive class (false negatives) as shown in Figure 18.

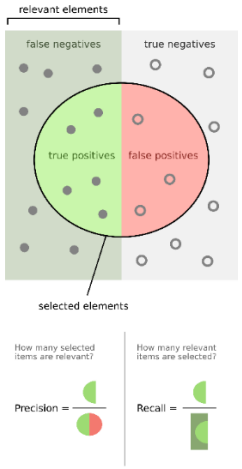
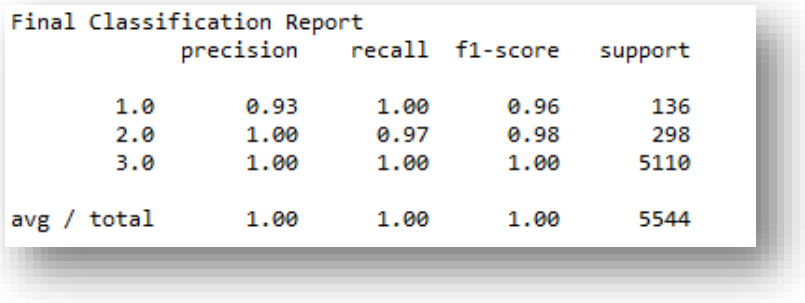


Figure 18: Precision and recall for class label

F1-score (also called F-measure) considers both the precision and the recall of the test to compute the score, it can be defined as a weighted average of the precision and recall, an F1 score reaches its best value at one and worst score at 0. The scores corresponding to every class to interpret the accuracy of the classifier in that particular class compared to all other classes. Finally, the support is the number of samples of the true response that lie in that class [25].

Figure 19 shows Classification report for training dataset, precision, from the confusion matrix of the training data set Recall, F1 score and support can be understandable.



Final Classification Report

	precision	recall	f1-score	support
1.0	0.93	1.00	0.96	136
2.0	1.00	0.97	0.98	298
3.0	1.00	1.00	1.00	5110
avg / total	1.00	1.00	1.00	5544

Figure 19: Classification report for training dataset

In regards to evaluation the classifier based on confusion matrix and calculating the error confusion matrix for the testing dataset 1656 instances with 99.82 % accuracy has been presented in details.

```
=== Confusion Matrix ===

    a    b    c   <-- classified as
[[  30    0    0]     a = 1.
 [   0   67    3]     b = 2.
 [   0    0 1556]]    c = 3.
```

In class (1) A there are 30 instances, 30 of them are correctly classified and zero are incorrectly classified. In class (2) B there are 70 instances, 67 of them are correctly classified and three are incorrectly classified. In class (3) C there are 1556 instances, 1556 of them are correctly classified and zero are incorrectly classified.

As a result: Sensitivity, which are true positive or correctly classified instances equal 1653 about 99.8188%. Whereas, Specificity, which are true negative or incorrectly, classified instances are three about 0.002 %

5.2.3 Tree visualization

Decision trees are powerful tools at the same time they are simple to utilize for visual analysis as well as knowledge ectraction. However, when applied to the dataset available nowadays which are complicated with huge amount of data, then decision trees tend to be complix, large, uneasy to visualize and difficult to understand. Even though this difficulty can be dominated by clustering the dataset and representing (visualizing) the decision tree of each cluster independently, in this thesis clustring has not been implemented, as the consentrating was on achieveing a good accuracy in order to apply the classification more efficiently for our PyDev project . A snapshot of our decision tree with our dataset after feature selection is shown in Figure 20 since the tree of the whole dataset is so complicated.

According to the figure below, ‘TSH’ attribute has been sat as the ‘root node’ which determine that 6489 instances belongs to class 3 ‘hypothyroidism’ due to having a value of TSH less than or equal to 0.006 . Whereas the rest of the samples which have a value of ‘TSH’ more than 0.006 is being tested on the first internal node which is ’FTI’ which is consist with what we had explained before in medical description section (Combining the TSH test with the FTI precisely determines how the thyroid gland is functioning). On

the same method, the tree is splitting until reaching the leaf node for each attribute. In this tree, there are 18 leaf node and 17 non-leaf node.

Based on our accuracy of 99.82% Percent, this is a very good model.

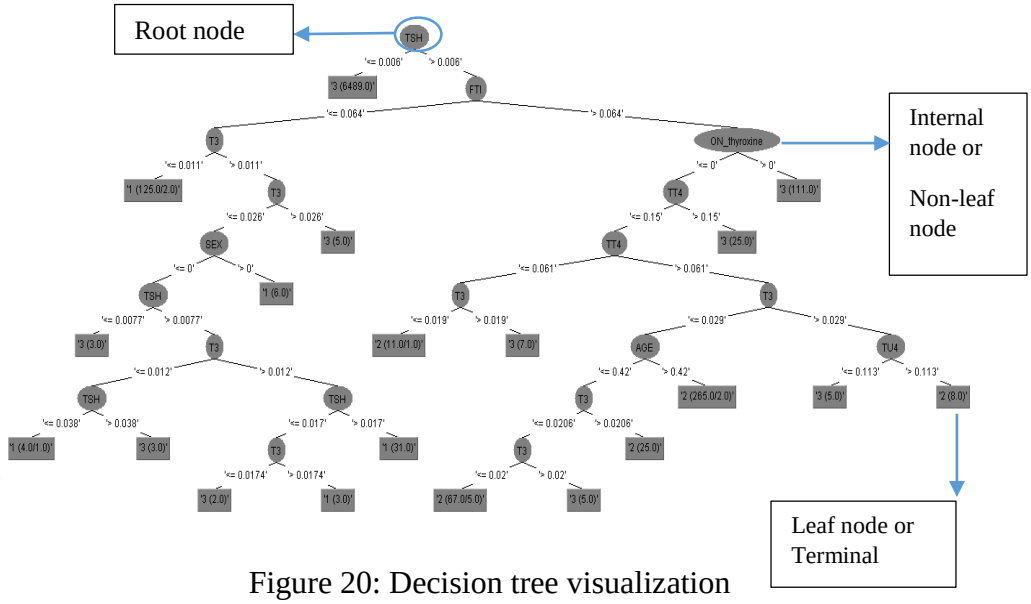


Figure 20: Decision tree visualization

5.3 Designing the Graphical user interface (GUI)

In computer science, a graphical user interface abbreviated GUI can be defined as a program interface that takes feature of the computer's graphics capabilities to create a program easy to use. Well-designed graphical user interfaces can make the user free from learning complicated command languages. In other words, GUI is a type of interface that enables users to be more interactive with electronic devices using graphical icons and visual indicators [33].

5.3.1 PyQt based application and QT designer

Qt is s a cross-platform application framework that widely used in the field of developing application software, in addition to its ability to run on various software and hardware platforms with small or sometimes no change in the underlying codebase, however it is still a native application with the and speed and capabilities.

PyQt is one of Python's options for GUI programming by being a Python binding of the cross-platform GUI toolkit Qt. PyQt is implemented as a Python plug-in. It is available to Qt versions older than 4.5. Both Qt, PyQt is free software. In the current work, PyQt4.1 has been used.

Qt designer is a C++ toolkit for cross plat form application development, it utilizes for building and designing graphical user interfaces (GUIs) from Qt components. As well as enabling the user to shape and customize dialogs or widgets in a what-you-see-is-what-you-get (WYSIWYG) manner, also to test them using several resolutions. Widgets and forms created with Qt Designer integrated easily into programmed code, by utilizing Qt's signals and slots mechanism, which lets the user to easily assign behaviour to graphical elements. Features such as custom plugins and widget promotion allow the user to use his/her own components with Qt Designer. Furthermore, all properties set in Qt Designer can be modified dynamically within the code [34] .

Qt Designer is used to create user interfaces for several purposes, and it provides various kinds of form templates for each graphical user interface. To create application windows with toolbars, menu bars, and many others the template ‘main window’ has been used.

5.3.2 Initial design

During training CART algorithm and the first design of MLTDD app, the study shows how it is important to have a complete dataset during designing process or during applying the algorithm, since 99.82% accuracy has been obtained by using all dataset's features. Therefore, in this work an initial GUI has been designed using Qt designer version 5.5.1, with the whole 21 features of the data set as shown in figure 21.

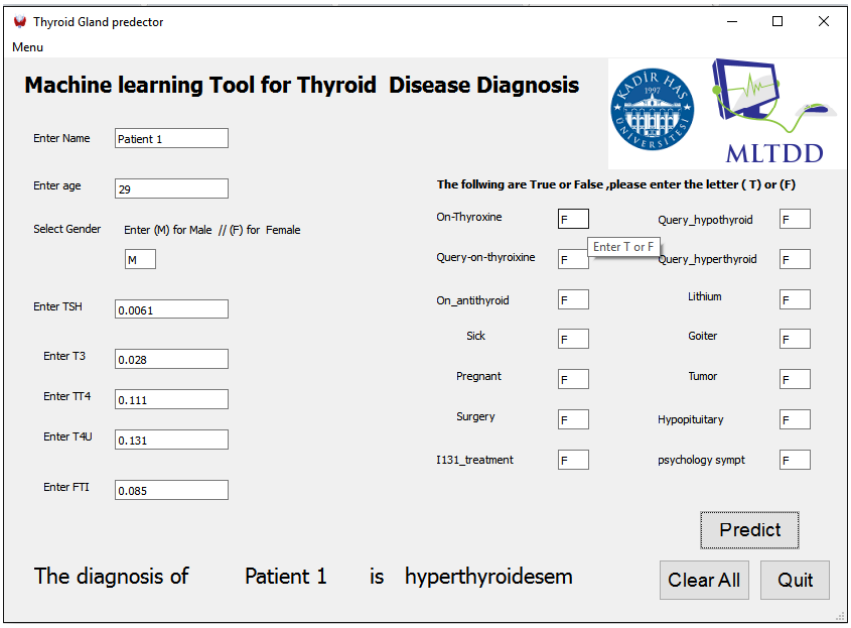


Figure 21: Snapshot of the initial design of MLTDD

5.3.3 After pre-processing

Feature Selection Techniques Used in Current Work is Information Gain, The information gain method was proposed to approximate quality of each attribute using the entropy by estimating the difference between the prior entropy and the post entropy as shown in Figure 22.

Eliminating the ten less important features which are the feature number 4query_on_thyroxine), 5 (on_antithyroid_medication), 6 (sick), 8 (Thyroid_surgery), 9 (131-treatment), 11 (query_hyperthyriod), 12 (lithium), 13 (goiter), 14 (tumor), 15 (hypopituitary), has been useful in terms of making our GUI more acceptable and easy to use as shown in Figure 23.

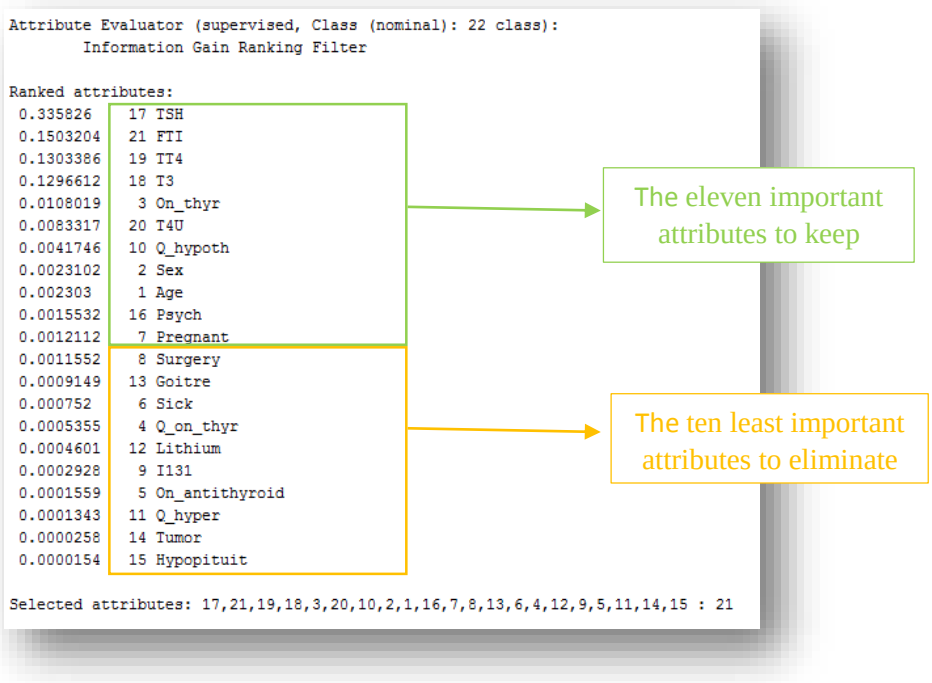


Figure 22: Information Gain Ranking Filter

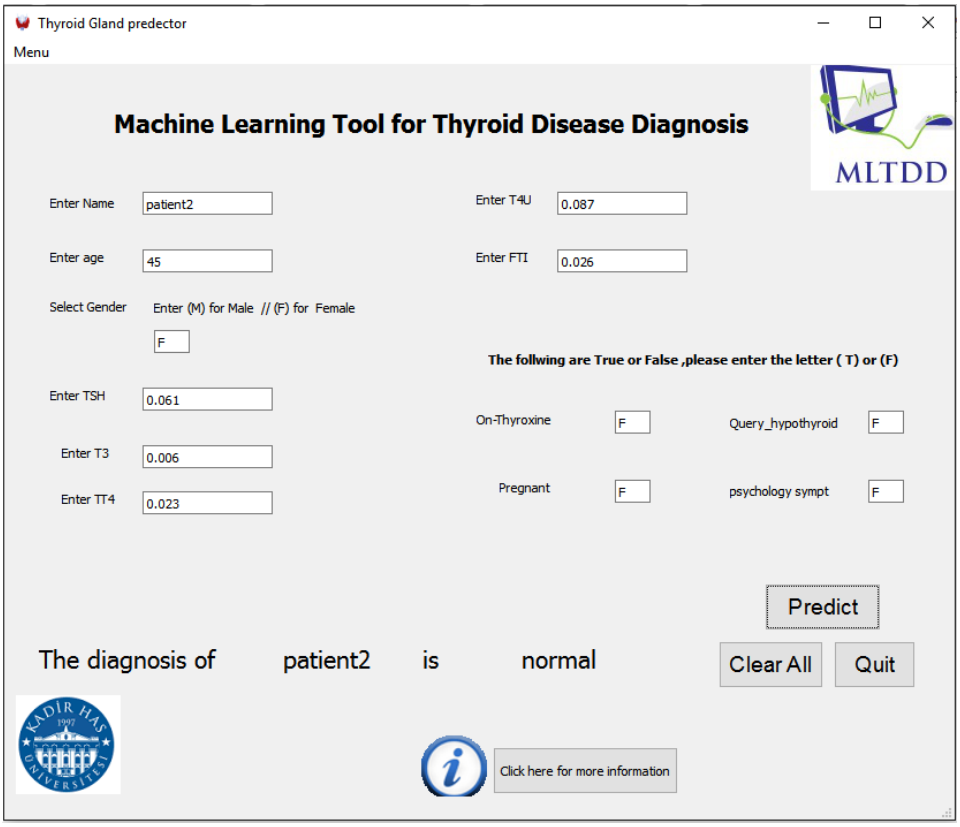


Figure 23: Snapshot MLTDD interface

5.4 Using Eclipse (PyDev) to connect GUI with python files

Eclipse can be defined as an integrated development environment (IDE) used in computer programming. It has a base workspace and plug-in system for customizing the environment. Eclipse is written mostly in Java and its main use is for developing Java applications, but it can also be used , through the use of plugins, to develop applications in other programming languages such as: C, C++, Fortran, JavaScript, Julia, PHP, Prolog, Python, and many others languages.

The Eclipse software development kit (SDK) is free and open-source software contains the Java development tools. It is fundamentally purposed for Java developers, but users may extend its abilities by installing others plug-ins written for the Eclipse Platform, like development toolkits for another programming languages, and then user can be able to write and contribute any plug-in modules. The initial codebase originated from IBM VisualAge [25].

Eclipse (Juno) version 4.2.2 also called Eclipse classic or eclipse SDK has been used in this works although there are many other versions such as Eclipse Kepler (4.3), Eclipse Luna (4.4), Eclipse Mars (4.5), and Eclipse Neon (4.6).

PyDev is a powerful plugin that enables Eclipse to be used as a Python IDE; also (Jython) and (IronPython) are supported. It uses an advanced type of inference techniques to add features such as code analysis and code completion, while still providing many others such as refactoring, a debugger, interactive console, etc. It is a nice IDE for python developer, which makes the world of python development a better place. It is critical that (java), (python) and (Eclipse) are all either 32 bit or 64 bit.

Building our project with PyDev has been done by importing our GUI from Qt Designer and connecting the 'predict' button with the python program, which has the task of training CART algorithm and generating a predictor model. Important programming steps has been done since our job as programmer is to enable users to enter the data in a normal format by converting some of them to other type of data in order to suit the predictor model. Figure 24 shows the first step has been done, which is to convert the gender data to a numeric data of a value 1 for male and zero for female, which is compatible with the origin dataset. The user is supposed enter 'F' for female and 'M' for male, which is more acceptable than entering zero for male and one for female (as a small 'tool tip' will appear for this instruction as in Figure 21) .

```
if self.sexvalu.text() == 'M' or self.sexvalu.text() == 'm'  :
        fg = 0
else:
        fg = 1

if self.ontvalu.text() == 'F' or self.ontvalu.text() == 'f'  :
        fh = 0
else:
        fh = 1
```

Figure 24: converting gender data to numeric data

In the origin data set the age is normalized into the decimal range 0.00-1.00, therefore, our job is to make it simple for users by enabling them to enter it as a normal age and then the program will convert it into normalized form in order to be suitable for the algorithm as in Figure 25.

```
def predbutton_clicked(self):
    nm = self.lineEdit.text()
    faa = int(self.agevalu.text())
    fa = float(faa / 100.0)

    fb = float(self.tshvalu.text())
    fc = float(self.ttrvalu.text())
```

Figure 25: Converting the age data

The next stage is to connecting the ‘predict’ button with the python program whose main porpos is apply and train CART algorithm using training dataset and then to create a predictor model as shown in Figure 26.

```
from sklearn.tree import DecisionTreeClassifier
dt = DecisionTreeClassifier(criterion='entropy', splitter='best', max_depth=None, min_samples_split=20,
                            min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None,
                            random_state=None, max_leaf_nodes=None, class_weight=None, presort=False)

dt.fit(X_train, y_train)
```

Figure 26: Apply and train CART algorithm

Setting some button actions has been done as shown in Figure 27, these button have a useful task such as predict button, clear button, quit button and info button which has the action of presenting some primary information about the prediction process.

```
self.predbutton.clicked.connect(self.predbutton_clicked)
self.actionPredict.triggered.connect(self.predbutton_clicked)


self.clearbutton.clicked.connect(self.clearbutton_clicked)
self.actionClear_All.triggered.connect(self.clearbutton_clicked)

self.quitbutton.clicked.connect(self.quitbutton_clicked)
self.actionExit.triggered.connect(self.quitbutton_clicked)

self.infobutton.clicked.connect(self.infobutton_clicked)
self.actionMore_info.triggered.connect(self.infobutton_clicked)
```

Figure 27: Setting button actions

5.5 Performance Results

According the investigation on machine learning algorithms, features selections techniques, and inference techniques this work proposed a prediction tool for diagnosing thyroid gland based on the best performed machine learning algorithm and the best performed feature selections methods with the interactive inference techniques. Therefore, we proposed a new tool based on decision tree algorithm and eclipse SDK. The predictive tool showed a classification accuracy of 99.82 %, which underline the tool capability on ann_thyroid Dataset.

The research introduced the term of classification on a well-known machine learning classifiers on thyroid gland datasets. Classification becomes a hot topic in the field of machine learning due the capability of reducing complex information and providing a guide for treatment and prognosis, assist in streamlining data collection, and assist clinicians to compare efficacy and efficiency of treatments. This study confirms that classification algorithms can improve the prediction accuracy. In addition, there is no best combination of classifiers that suites all datasets. However, no specific features selections method that suits all machine-learning tools.

Based on the experiments on different machine learning algorithms and Ann-Thyroid Dataset, the study may conclude that CART algorithm performed better accuracy comparing with two well-known algorithms. Table 4 shows that by using support vector machine algorithm (SVM) the best accuracy has been obtained as follows. Prediction accuracy for training dataset 92.45% Prediction accuracy for testing dataset 93.06%. Using k-nearest neighbor algorithm (KNN) the best accuracy has been obtained as follows. Prediction accuracy for training dataset 94 % whereas Prediction accuracy for testing dataset 94.82% .With 8 nearest neighbor which has been selected using GridSearchCV function that does all the hard work and returns the best k parameter .

Algorithm	Accuracy (%) Training dataset	Accuracy (%) Testing dataset
Decision tree CART	99.62 %	99.82 %
Support vector machine SVM	92.45 %	93.45 %
k-nearest neighbor algorithm KNN	94 %	94.82 %

Table 4: Accuracy comparison for three algorithms

Among various preprocessing methods, in this study we applied firstly principle component analysis (PCA) for feature selection but that the accuracy decreased from 99.644 to 97.533, secondly, we applied another preprocessing method called (ROUNDUM Subset) but we got the same result. For the purpose of decreasing the number of features to be able to design an acceptable GUI, since the dataset has 21 features information gain attribute evaluation has been applied as a feature selection method. A good ranking has been obtained which helps to eliminate the ten least important attributes and keep the other 11 attributes. With 11 features, we got a 99.70 % accuracy, which is not much decreasing, since the best accuracy has been achieved with the all twenty-one features is 99.82% with 23% testing dataset as shown at Table 3. Therefore, we decided to eliminate those 10 feature to make our tool interface more acceptable and easier to use. In Table 5, there are different accuracy values with different splitting of dataset comparing before and after applying the feature selection method and eliminating the less important features, in Figure 28 the graph shows this comparison as well.

Percentage split for testing dataset	Accuracy (%) With 21 features	Accuracy (%) With 11 features
10 %	99.58 %	99.17 %
15 %	99.72 %	99.35 %
20 %	99.79 %	99.51 %
23 %	99.82 %	99.58 %
30 %	99.68 %	99.68 %
32 %	99.65 %	99.70 %
33 %	99.62 %	99.66 %

Table 5: Specifies the values of Accuracy before and after eliminating 10 features.

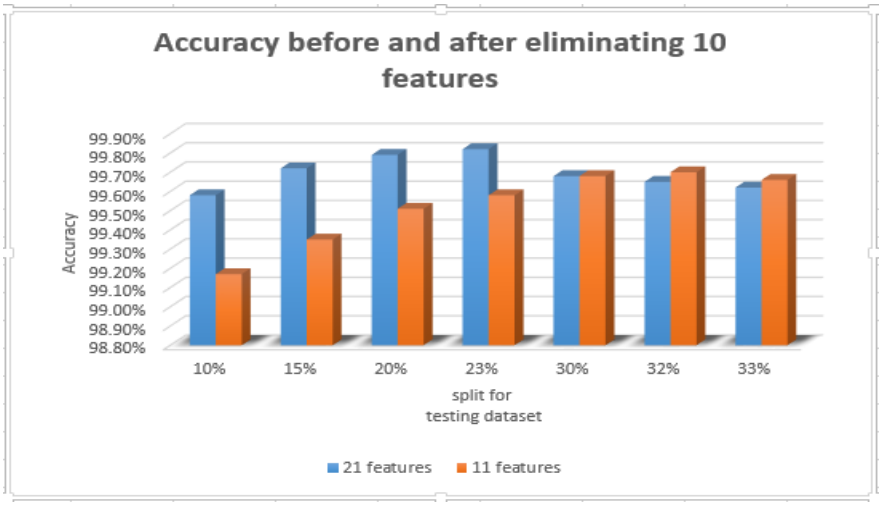


Figure 28: Accuracy comparison graph for the splitting percentage

Chapter 6

Discussion and Related work

The main purpose of current research is to participate in the efforts of enhancing the quality of healthcare services, proposing technology as one of solutions for the problem of diagnosis. However, the process of utilizing technology in healthcare services is a comprehensive process and involves many stages and steps. It is very important to discuss all related issues to conclude with new system that derive the expected services.

The combination created a new app for diagnosing thyroid gland by reducing the number of features to the optimal number using the information gain and then applied the CART algorithm to the dataset after designing a GUI to create a predictor tool. The study found that the accuracy for the proposed tool is 99.82% compared with other methods. The proposed tool showed a very promising results which may lead to further attempts to utilize information technology for diagnosing patients for thyroid gland disease.

Another important issue when dealing with databases and health databases is features selections techniques and how to determine the most important features that lead to more accurate diagnosis. In general, features selections methods can improve the performance of learning algorithms, According to the results obtained by the current work on ann-thyroid dataset; CART has performed the supreme concerning classification accuracy. Although CART have performed just better on dataset after applying features selections methods than the original dataset with no features selections techniques. Therefore, machine-learning researcher should understand the nature of datasets and learning algorithm characteristics in order to obtain better outcomes as possible.

In regards to improving our app as a future work we aim to convert our PyDev project into a web project and for this reason, we have changed our work from Spyder to Eclipse.

However, Spyder has a direct link to Qt Designer but our project has been changed to Eclipse since it enables users to transform a python project to web project and unleash their app on the world by deploying it to App Engine. Google App Engine (GAE) is plugged in Eclipse and can be used by using the option (install new software), which enables installing google app engine in eclipse SDK.

In addition, Thyroid dataset used in this study has only three possible class label. Clinical practice; however, is often more complex and outcomes maybe in different format. It is envisaged that the future work can contribute to the knowledge and improve the accuracy and reliability of established tool by broaden the databases and expanding the criteria for measuring the performance of established tool.

Related work

This part of research contains a comparison of our work with the existing works which had the same goal of thyroid gland diagnosis.

Anupam Skukla et al. [35] suggested the diagnosis of thyroid disorders with Artificial Neural Networks (ANNs). Some ANN algorithms have been utilized for the diagnosis, which are the Learning Vector Quantization (LVQ) Networks, the Radial Basis Function (RBFN), and the Back propagation algorithm (BPA).

Lale Ozyilmaz et al. [36] focused on suitable interpretation of the thyroid data beside complementary investigation and the clinical examination as a critical issue on the diagnosis of thyroid disease. Several neural network methods have been implemented for the diagnosis of thyroid disorder, which are adaptive Conic Section Function Neural Network (CSFNN), back-propagation (BP), fast back-propagation (FBP), and Multi-Layer Perceptron (MLP).

Fatemeh Saiti et al. [37] proposed two algorithms, which are Support Vector Machines and Probabilistic Neural Network considering separating and classification the Hypothyroidism and hyperthyroidism diseases, which plays a vital role for thyroid diagnosis. These methods depend on robust classification algorithms, in order to deal with irrelevant and redundant features. Genetic Algorithm is a really optimization technique. As it operates iteratively on a population of structures each of them represent a candidate solution to the problem duly encoded as a string of symbols (e.g. Binary). Three basal genetic operators guide the search: mutation, selection, and crossover. The PNN is a supervised neural network and it is widely used in the area of nonlinear mapping, pattern recognition, and likelihood ratios.

Carlos Ordonez et al. [38] compared decision tree rules with association rules. Association rules research for hidden patterns turning them out to be suitable for discovering predictive rules including subsets of the medical data set. As result, association rules are unrelated, appear at low support and discovered rules are large in number. Whereas the decision tree represents a predictive model of the data set. The large number of discovered association rules may obtain rule summarization. A decision tree could have at least 50% prediction accuracy and generally over 80% accuracy for binary target variables.

A.S.Varde et al. [39] has developed the clinical laboratory expert system in 1991 for the diagnosis of thyroid disorder. The system had considered clinical findings and the results of applicable laboratory tests along with the patient's medical history. The system had been execute using VP-Expert, version 2.02 that is commercially available software.

Palanichamy Jaganathan et al. [40] developed F-score feature selection method that used to nominate the most relevant features for classification of thyroid dataset. The average progressed F-score value of the features computed is set as the severance for electing the

features. The features above the cut-off are picking as more relevant than others. Then the chosen features were used in thyroid disease diagnosis with and C4.5 algorithm and multilayer perceptron (MLP), then compared with recent research results obtained from generalized discriminant analysis-wavelet support vector machine (GDA-WSVM) technique (91.86%). The result shows that their new feature selection method applied to this dataset has generated better classification accuracy than GDA-WSVM combination, with an amelioration of 1.63% of accuracy for improved F-score-MLP combination (93.49%).

Conclusion

In recent years, an enormous development in medical expert systems has been seen, and the systems now available are grown enough for targeted adoption in practice. However, expert systems can be progressively integrated in hospital information systems in order to deliver health-care even more effectively.

Overall thyroid disease can be tricky to diagnose because symptoms are easily confused with other illness condition. When thyroid disease is caught early, treatment can control the dysfunctionality. In this study, diagnosing thyroid disease is aimed with a machine-learning tool called as MLTDD (machine learning tool for thyroid disease diagnosis). MLTDD is a tool, which is presented to endocrinologists or students, could predict diagnosis with 99.7% accuracy for thyroid diseases. According the investigation, using this tool in term of diagnosis process provides a more colorful and variegated environment for the users than huge and hard-covered materials. However, students studying endocrinology for thyroid diseases can use this tool for testing their knowledge by comparing their predictions with MLTDD.

Future work can be described as follows. The current research resided mainly on classification accuracy as the main criteria for measuring the performance of proposed tool. However, future work will focus in other criteria such as classification speed and computational cost.

References

[1] Gunter-TD, and, Terry-NP, the Emergence of National Electronic Health Record Architectures in the United States and Australia: Models, Costs, and Questions. J Med Internet Res. 2005.

[2] Michael B. Zimmermann, Iodine deficiency. Endocr Rev 2009. Available: http://press.endocrine.org/doi/abs/10.1210/er.2009-0011

[3] Mark P. J. Vanderpump and W. Michael and G. Tunbridge, The epidemiology of thyroid diseases. In: Braverman LE,Utiger RD, editors. Werner and Ingbar's The Thyroid: A Fundamental and Clinical Text. Ninth edn. Philadelphia: JB Lippincott-Raven; 2005.

[4] Tunbridge WMG, Evered DC, Hall R, et al. The spectrum of thyroid disease in the community, 1977.

[5] Vander JB, Gaston EA, Dawber TR. The significance of nontoxic thyroid nodules: final report of a 15-year study of the incidence of malignancy. Ann Intern Med 1968.

[6] Reiners C, Wegscheider K, Schicha H, et al. Prevalence of thyroid disorders in the working population of Germany: ultrasonography screening in 96,278 unselected employees. Thyroid 2004.

[7] Perros P, editor; British Thyroid Association and Royal College of Physicians. Report of the Thyroid Cancer Guidelines Update Group. London: Royal College of Physicians; 2007. Guidelines for the management of thyroid cancer.http://www.british-thyroid-association.org/news/Docs/Thyroid_cancer_guidelines_2007.pdf.

[8] Davies L, Welch HG. Increasing incidence of thyroid cancer in the United States, 1973–2002. JAMA 2006.

[9] International agency for research on cancer, world health organization

GLOBOCAN 2012 estimate cancer incidence , mortality and prevalence worldwide in 2012. http://globocan.iarc.fr/Pages/fact_sheets_population.aspx

[10] Grüters A, Krude H, on the management of congenital hypothyroidism. 2007.

[11] Right diagnosis for health grades. Web site: http://www.rightdiagnosis.com/t/thyroid/stats-country.htm#extrapwarning .

[12] Rokach, L. and O. Maimon, eds. Data Mining and Knowledge Discovery, 2010.

[13] Tarca, A.L., et al., Machine Learning and Its Applications to Biology. PLoS Comput Biol, 2007.

[14] Machine Learning Stanford University, https://www.coursera.org/learn/machine-learning, Instructors,Andrew Ng,Associate Professor, Stanford University; Chief Scientist, Baidu; Chairman and Co-founder, Coursera

[15] ForbesWoman, The Little Black Book of Billionaire Secrets, JAN 6, 2014 @ 02:45 PM 13,355 VIEWS, Six Novel Machine Learning Applications

[16] Venture radar, Discover the Innovators and Disruptors, VentureRadar discovers and ranks companies, making them visible to potential partners, customers, and investors. Website: http://blog.ventureradar.com/2015/09/30/top-5-companies-revolutionizing-healthcare-with-machine-learning/

[17] The editor-in-chief at MachineLearningMastery.com. Jason is a husband, father, researcher, author, professional programmer and a machine learning practitioner. Website: http://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

[18]Rokach, L., Data mining with decision trees: theory and applications. Vol. 69. 2007: World scientific.

[19] Tan, P.-N., M. Steinbach, and V. Kumar, Introduction to Data Mining. 2006: Addison-Wesley.

[20] F.V. Nowak, Ohio University, The Thyroid Gland: Function and Regulation, 2009.

[21] Neepa Y. Choksi, Gloria D. Jahnke, Cathy St. Hilaire, and Michael Shelby, Role of Thyroid Hormones in Human and Laboratory Animal Reproductive Health, Published 2003 Wiley-Liss, Inc.

[22] Medicine net. Website: http://www.medicinenet.com/goiter_causes_symptoms_and_treatment/views.

[23] Machine learning databases. Web site: http://archive.ics.uci.edu/ml/machine-learning-databases/ thyroid-disease/

[24] Endocrinology An Integrated Approach, Stephen Nussey and Saffron Whitehead. St. George's Hospital Medical School, London, UK Oxford: BIOS Scientific Publishers; 2001. ISBN-10: 1-85996-252-1

[25] Wikipedia. Web site: https://en.wikipedia.org

[26] Dan A. Simovici, Szymon Jaroszewicz, A Metric Approach to Building Decision Trees Based on Goodman-Kruskal Association Index. http://link.springer.com/chapter/10.1007%2F978-3-540-24775-3_23.

[27]Robert Nisbet (Author), Gary Miner (Author), John Elder IV (Author), Handbook of Statistical Analysis and Data Mining Applications 1st Edition.

[28]H.S.Hota, Diagnosis of Breast Cancer Using Intelligent Techniques, International Journal of Emerging Science and Engineering (IJESE), January 2013

[29] Jiawei Han, Micheline Kamber, Data Mining Concepts and Techniques. Published by Elsevier 2006.

[30] Michael Bowles, Machine Learning in Python: Essential Techniques for Predictive Analysis, 2015

[31] David Mertz, Functional Programming in Python, 2015.

[32] Spyder 2.3 documentation. Web site: https://pythonhosted.org/spyder/

[33] Graphical User Interface. Web site: http://www.webopedia.com/

[34] Qt Designer Manual. Web site: http://doc.qt.io/qt-4.8/designer-manual.html

[35] Anupam Shukla, Prabhdeep Kaur, Ritu Tiwari and R.R. Janghel, Diagnosis of Thyroid disease using Artificial Neural Network. In Proceedings of IEEE IACC 2009.

[36] Lale Ozyilmaz , Tulay Yildirim, Diagnosis of Thyroid disease using Artificial Neural Network Methods. In Proceedings of ICONIP 2002.

[37] Fatemeh Saiti and Mahdi Aliyari, Thyroid Disease Diagnosis based on Genetic algorithms using PNN and SVM. In Proceedings of IEEE 2009.

[38] Carlos Ordonez University of Houston, Houston, TX, "Comparing association rules and decision trees for disease prediction". In Proceedings of Int. Conf. Inf. Knowl. Manage, 2006.

[39] A.S.Varde,K.L.Massey and H.C.Wood, "A Clinical Laboratory Expert System for the Diagnosis of Thyroid Disfunction", Proceeding of Annual International Conference of the IEEE Engineering in Medicine and Biology Society.

[40] Palanichamy Jaganathan and Nallamuthu Rajkumar, An expert system for optimizing thyroid disease diagnosis. In Proceedings of International Journal of Computational Science and Engineering, 2012.

Curriculum Vitae

Izdihar Mofek was born in October 27th, 1985, in Misurata-libya. She received her Higher Diploma in Computer Technology in 2005 from higher comprehensive professions institute-Libya. Since 2005 until 2012, she worked as a demonstrator, higher comprehensive professions institute.