

# MLTDD : USE OF MACHINE LEARNING TECHNIQUES FOR DIAGNOSIS OF THYROID GLAND DISORDER

Izdiyar Al-muwaffaq and Zeki Bozkus

Department of Computer Engineering, Kadir Has University  
izdiyar.Mofek@stu.khas.edu.tr, zeki.bozkus@khas.edu.tr

## ABSTRACT

*Machine learning algorithms are used to diagnosis for many diseases after very important improvements of classification algorithms as well as having large data sets and high performing computational units. All of these increased the accuracy of these methods. The diagnosis of thyroid gland disorders is one of the application for important classification problem. This study majorly focuses on thyroid gland medical diseases caused by underactive or overactive thyroid glands. The dataset used for the study was taken from UCI repository. Classification of this thyroid disease dataset was a considerable task using decision tree algorithm. The overall prediction accuracy is 100% for training and in range between 98.7% and 99.8% for testing. In this study, we developed the Machine Learning tool for Thyroid Disease Diagnosis (MLTDD), an Intelligent thyroid gland disease prediction tool in Python, which can effectively help to make the right decision, has been designed using PyDev, which is python IDE for Eclipse.*

## KEYWORDS

*Machine Learning, Thyroid diseases, CRT decision tree algorithm, PyDev, Python IDE.*

## 1. INTRODUCTION

Classification algorithms are very important category of supervised machine learning algorithms. These algorithms require a very large training sets. These training data sets are consisting of many features or attributes which describe the individual sample. Since we are doing supervised learning algorithm. All of the training set are labelled correctly. The classification algorithms such as decision trees and support vector machines (SVM), develop model with these data with many different parameters. When we have a new unlabeled sample, we can use the model to predict the label of the new sample. These techniques are used for disease diagnosis to help doctor to effectively label the new case.

The thyroid releases two principal hormones. The first is called thyroxine (T4) and the other one is triiodothyronine (T3) into the blood stream. The main functions of the thyroid hormones are to regulate the growth rate of metabolism. There are two common problems of thyroid disorder: Hyperthyroidism and Hypothyroidism. The first one releases too much thyroid hormone into the blood stream and the second one releases too low thyroid hormones to the blood stream. These

means that thyroids are very active in Hyperthyroidism. In contrast, thyroids are not active in Hypothyroidism [1] [2].

A decision tree is one of the very effective machine learning classifier where the algorithm makes a tree structure, where every non-leaf node denotes a test on an attribute, each branch performs an outcome of the test and each leaf node holds a class label. Decision tree is very simple and effective classifier so this algorithm can be used in several application areas such as medicine, financial analysis, astronomy and molecular biology for classification [3] [4]. We are going to use decision tree to identify and predict thyroid diseases whether new patient data (symptom or features) show thyroid disorder or normal.

This study developed a tool to be used for diagnosing thyroid diseases. We named our tool as a MLTDD (machine learning tool for thyroid disease diagnosis). MLTDD could predict 99.81% accuracy thyroid diseases with sample dataset.

The paper is organized as follows: The Section 2 describes the thyroid dataset and introduces decision tree techniques. The obtained experimental results in application are given in Section 3. Discussions and comparisons with previous work will be found in Section 4. Finally, Section 5 presents the conclusions.

## 2. METHODOLOGY USED

This section explains about the algorithm, language and software used for this work. The data set used for experimental purpose is downloaded from university of California of Irvin (UCI) repository site (web source <http://www.archive.ics.uci.edu/ml/datasets.html>). Details of data set is given in Section 2.1.

An implementation process of thyroid gland diagnosis process is depicted in Figure 1.

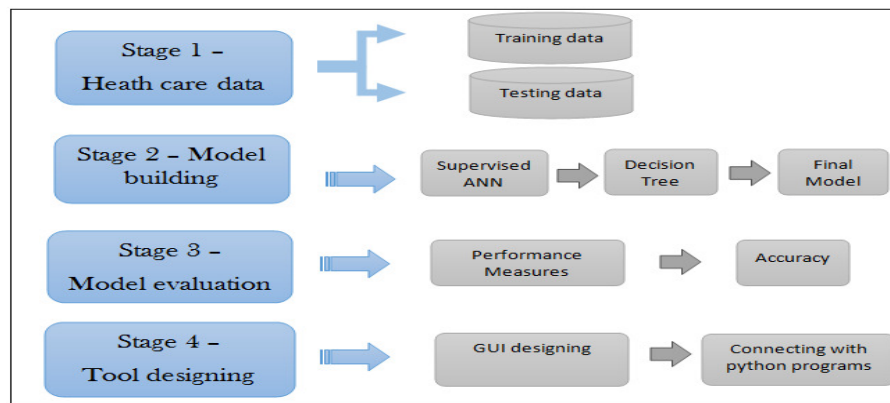


Figure 1: Implementation of MLTDD Process

### 2.1 Description of data-set

Thyroid dataset for this work had been collected from UCI machine learning repository. It consists of 7200 instances and 3 classes, 3772 are training instances, 3428 testing are instances

and 21 attributes as shown in Table 1 [5]. The task is to detect is a given patient has a normal condition (1) or suffers from hyperthyroidism (2) or hypothyroidism (3).

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. Three classes are assigned to each of the measurement vectors, which correspond with hyper-rthyroidism, hypothyroidism and normal function of the thyroid gland (Table 2).

Table 1. General information

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. Attribute description.

Attribute	Domain	Attribute	Domain	Attribute	Domain
Age	[0.01, 0.97]	Thyroid surgery	[0, 1]	Hypopituitary	[0, 1]
Sex	[0, 1]	Il31_treatment	[0, 1]	Psych	[0, 1]
On_thyroxine	[0, 1]	Query_hypothyroid	[0, 1]	TSH	[0.0, 0.53]
Query_on_thyroxine	[0, 1]	Query_hyperthyroid	[0, 1]	T3	[0.0005, 0.18]
On_antithyroid_medication	[0, 1]	Lithium	[0, 1]	TT4	[0.002, 0.6]
Sick	[0, 1]	Goiter	[0, 1]	T4U	[0.017, 0.233]
Pregnant	[0, 1]	Tumor	[0, 1]	FTI	[0.002, 0.642]
Class	{1,2,3}				

## 2.2 Decision Tree classification method

We used decision tree classification for our thyroid diseases diagnosis tool. Decision tree takes training datasets and construct an internal tree structure. The non-leaf node will have a question or a condition to check based on feature of the data. The leaf nodes of the tree have the labels.

The algorithm of constructing decision tree is a divide-and-conquer algorithm. The root of the tree asks one of the feature questions to splits datasets into multiple branches based on the answer of the parent question. If all the subsets of data in the branch have the same label, that branch will stop. It will not grow from that branch any further. Otherwise, the constructor algorithm can split the data further with new conditions from other features of the data sets. The algorithm will recursively try to create new branches of the tree based on the features (or attributes) of datasets. Basically algorithm try to understand which features effect the decision for labeling.

When we have a record of a new patient, we use the decision tree as if it is a flow-charts. We try to go the labeled leaf node by asking the same questions of the tree until we reach the leaf node. We then classify the new patient the same as the leaf node label which we reach from the root of the decision tree. There are huge number of research how to select which attributes is best to select for the root and non-leaf nodes. Basically, the constructor will choose the attributes to have maximum information gains [6].

## 2.3 MLTDD: Intelligent Thyroid Gland Disease Prediction System

We designed this application to be used by doctors, who are not expert computer users. So our design goal was to develop the user interface of the thyroid diseases application as user friendly as possible. The application has been implemented in Eclipse Python (PyDev plugged in eclipse JUNO Version: 4.2.2) coded in python 2.7. The GUI has been designed in Qt designer version 5.5.1 which is graphical user interface for Qt application. MLTDD runs on windows environment (Fig. 2).

Thyroid Gland predictor

Menu

**Intelligent Thyroid gland Disease Prediction System**

Enter Name:

Enter age:

Select Gender: Enter (M) for Male // (F) for Female

Enter TSH:

Enter T3:

Enter TT4:

Enter T4U:

Enter FTI:

The following are True or False ,please enter the letter ( T ) or ( F )

On-Thyroxine:

Pregnant:

Query\_hypothyroid:

psychology sympt:

**Predict**

The diagnosis of patient 1 is normal

**Clear All** **Quit**



  [Click here for more information](#) Prediction accuracy is 99.7 % >> Decision Tree algorithm CRT was used>>

Fig. 2. Screen of prediction test of MLTDD app.

Ann dataset has been used to train the decision tree algorithm and create a model then predict a diagnosis for a new patient whose data will be entered by the user. Taking the data from GUI and pass them to a python program which uses the model created to classify the patient according to its data as a hypothyroidism or hyperthyroidism or a normal diagnosis, then by passing the diagnosis to the program which dealing with the GUI so that it can be printed out to the user.

### 3. RESULTS

For preprocessing, we applied “PCA” for feature selection but we noticed that the accuracy decreased from 99.644 to 97.533, we also applied another preprocessing method “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 and it doesn't decrease so much, since the best accuracy we got with the all 21 features is 99.82% with 23% testing dataset as shown at Table 3. So we decided to eliminate those 10 feature to make my "GUI" more acceptable and easier to use. In Table 3, 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 3 the graph shows this comparison as well.

Table 3: Specifies the values of Accuracy before and after eliminating 10 features.

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 %

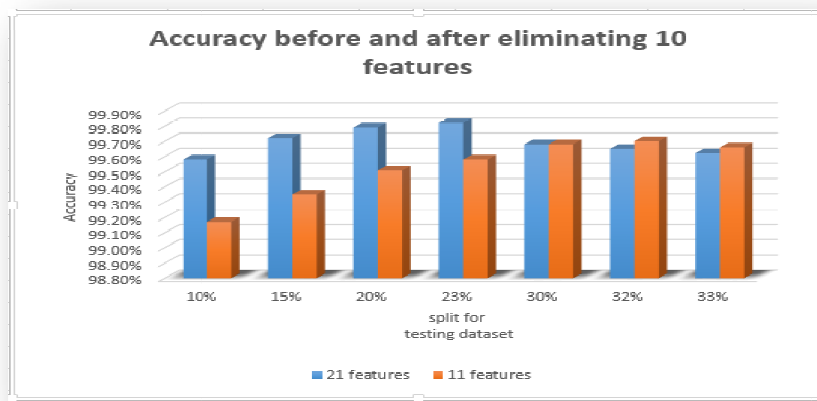


Figure 3: Accuracy comparison graph for the splitting percentage

## 4. RELATED WORK

Many methods and algorithms have been utilized in the past in medical disease classification. A skimpy abstract of them as follows:

Anupam Skukla et al. [8] suggested the diagnosis of thyroid disorders with Artificial Neural Networks (ANNs). Three ANN algorithms were; the Radial Basis Function (RBFN), the Back propagation algorithm (BPA), and the Learning Vector Quantization (LVQ) Networks have been utilized for the diagnosis.

Lale Ozyilmaz et al. [9] concentrated on suitable interpretation of the thyroid data. Several neural network methods such as fast back-propagation (FBP), Radial Basis Function (RBF), adaptive Conic Section Function Neural Network (CSFNN), and Multi-Layer Perceptron (MLP) with back-propagation (BP) have been utilized and compared for the diagnosis of thyroid disease.

Fatemeh Saiti et al. [10] 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.

Carlos Ordonez et al. [11] 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.

A.S.Varde et al. [12] 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. [13] developed F-score feature selection method that used to nominate the most relevant features for classification of thyroid dataset. 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%).

## 5. CONCLUSIONS

In recent years, machine learning is becoming very important computer science fields. It is slowly changing many industry including health industry. We collect many data and we have very fast computational power. Together with very effective algorithms, we can develop tools to help human to do their job better. In this study we develop a tool to help doctor for the disease diagnosis. We do not expect our tool to be used by a doctor as it is, but these kind of tool can be used by medical student to be a learning tool.

Thyroid disease can be tricky to diagnose because symptoms are easily confused with other illness condition. When thyroid disease is caught early, treatment can be given patients very effectively. In this study, diagnosing thyroid disease is aimed with a machine learning tool called

as MLTDD (machine learning tool for thyroid disease diagnosis). MLTDD could form a foresight diagnosis with 99.7% accuracy for thyroid diseases in the datasets we used.

In our opinion, we can develop some more diseases diagnosis tool to help medical students to learn the disease characteristics for testing purposes. We do not believe these tools will replace the doctor yet. But they can be a helper to doctor if we have further improvement at technology. However, students who is studying endocrinology for thyroid diseases can use this tool for testing their knowledge by comparing their predictions with MLTDD. We have to collect more data to training these algorithms. The more data means we need to have high performing computational units. We believe that when we put big data, high speed computational power and advance machine learning algorithm, we can have very effective diagnostic tools.

## REFERENCES

- [1] Dr. Sahni BS, Thyroid Disorders .Online Available:  
<http://www.homoeopathyclinic.com/articles/diseases/tyroid.pdf>
- [2] Thyroid: <https://en.wikipedia.org/wiki/Thyroid>
- [3] Jiawei Han , Micheline Kamber, Data Mining Concepts and Techniques. Published by Elsevier 2006.
- [4] V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman, Decision trees: An overview and their use in medicine. In Proceedings of Journal Medical System 2002.
- [5] <http://archive.ics.uci.edu/ml/machine-learning-databases/thyroid-disease/>
- [6] H.S.Hota, Diagnosis of Breast Cancer Using Intelligent Techniques, International Journal of Emerging Science and Engineering (IJESE), January 2013 .
- [7] Jiawei Han , Micheline Kamber, Data Mining Concepts and Techniques. Published by Elsevier 2006.
- [8] Anupam Shukla, Prabhdeep Kaur, Ritu Tiwari and R.R. Janghel, Diagnosis of Thyroid disease using Artificial Neural Network. In Proceedings of IEEE IACC 2009.
- [9] Lale Ozyilmaz , Tulay Yildirim, Diagnosis of Thyroid disease using Artificial Neural Network Methods. In Proceedings of ICONIP 2002.
- [10] Fatemeh Saiti and Mahdi Aliyari, Thyroid Disease Diagnosis based on Genetic algorithms using PNN and SVM. In Proceedings of IEEE 2009.
- [11] 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.
- [12] 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.
- [13] Palanichamy Jaganathan and Nallamuthu Rajkumar, ” An expert system for optimizing thyroid disease diagnosis”. In Proceedings of International Journal of Computational Science and Engineering,2012.