SJIF Impact Factor: 6.260| ISI I.F.Value:1.241 Volume: 4 | Issue: 9 | September | 2019 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)

# PERFORMANCE ANALYSIS OF THYROID HORMONAL DISEASE USING TREE-BASED ALGORITHMS

**[1] Dr Ambrose Akinbohun (MD)**
[1]*University Medical Sciences, Akure, Ondo State, Nigeria*

**[2]Folake Akinbohun**
[2]*Rufus Giwa Polytechnic, Owo, Ondo State, Nigeria*

**[3]Oyinloye Oghenerukevwe E**
[3] *Ekiti State University, Ado-Ekiti, Ekiti State, Nigeria*

**ABSTRACT**
*Thyroid hormones are biochemical substances that are essential for the metabolic activities of the body cells. The living cells of the human body depend on thyroid hormones to function properly. The thyroid hormones are produced by the thyroid gland. These hormones - thyroxine (T4), triiodothyroxine (T3) and thyroid stimulating hormone (TSH) must be produced in normal amounts (euthyroid state) for the body to function optimally. If the thyroid gland over-secretes hormones, it results in hyperthyroidism and if it does not produce up to what the body needs, it results in hypothyroidism. Therefore there is need to determine a tree-based predictive model that classifies the thyroid hormonal diseases. The objective of the study is to compare the performance analysis of the tree-based models, and pick the best model that is suitable for predicting thyroid hormonal diseases. The thyroid dataset from UCI machine learning repository was used for the tree-based models. The dataset was divided into training data and test data. Three tree-based models were adopted such as REPTree, Decision tree (C4.5) and Random forest. The models were evaluated using k cross validation on the dataset where k = 10. The results showed that Random Forest had highest accuracy and highest F1 score compared to REPTree and Decision Tree (C4.5). Random Forest is a good predictive model for thyroid hormonal diseases.*

**KEYWORDS:** *thyroid stimulating hormone, REPTree, Hypothyroidism, C4.5*

## 1.0 INTRODUCTION

Thyroid hormones are essential for regulation of energy metabolism, growth, neuronal development and reproduction. Hypothyroidism and hyperthyroidism are common conditions with potentially devastating health consequences that affect all populations worldwide [1]. Iodine nutrition is a key determinant of thyroid disease risk; however, other factors such as ageing, smoking status, genetic susceptibility etc influence thyroid disease epidemiology.

Medical field is constantly looking for more methodologies to keep its upward trend in predictive diagnosis. The medical records of patients have led to the storage of amounts of data. The data comes from various hospitals based on historical patients' records. This type of information has been stored in data-warehouses and mostly used to predict the category of disease a patient is likely to have. Predictive

algorithms and machine learning can give a better predictive model of any disease that doctors can use to educate patients.

As machine learning becomes more accessible, it plays more important role in predictive applications. Machine learning is the technology or tools that clinicians use to improve ongoing healthcare in medicine. If a physician sees a patient and enters symptoms and signs; the result of the data is produced. In this case, there is machine learning behind the scenes reading everything about that patient, and prompting the doctor with useful information for making a diagnosis. It has capabilities that provide value from a specific technological application in healthcare, and then doctors must take an incremental approach. Corbett (2018) [2] opined that machine learning needs a certain amount of data to generate an effective algorithm. Much of machine learning initially comes from organizations with big datasets. It is enabling comparative effectiveness, research, and producing unique, powerful machine learning algorithms.

Thyroid disease is the predominant form of thyroid dysfunction in the developed world. When the gland produces too much thyroid hormone, the condition is known as hyperthyroidism. When there is too little thyroid hormone produced, it is called hypothyroidism. These thyroid disorders can be predicted in a patient using predictive science. In Predictive science, data mining techniques extracts potentially useful information from databases which can be used for clinical decision-making to predict whether a patient has thyroid hormone falls under hyperthyroidism, hypothyroidism and euthyroidism (normal).

## 2.0 LITERATURE REVIEW

This section is divided into sections reviewing researches on thyroid disease and thyroid disease using learning algorithms.

### A.Related work on thyroid disease

Taylor *et al.* [1] researched on global epidemiology of hyperthyroidism and hypothyroidism. The authors reviewed the global incidence and prevalence of hyperthyroidism and hypothyroidism.

Ogbera *et al.* [3] wrote a paper on pattern of thyroid disorders in the southwestern region of Nigeria. The study attempted to describe the patterns of thyroid disorders, clinical features, and complications as seen in Nigerians. It showed that there was an upsurge in reported cases of thyroid disorders in Africans from 1970s. No model was used in the work.

Bemben *et al.*, [5] showed that among 283 people aged 60 years and older attending a primary care geriatrics clinic in Oklahoma, the prevalence of subclinical hypothyroidism was 15 percent and that of overt hypothyroidism was 1 percent in both women and men

[6] was of the opinion that incidence of well-differentiated thyroid cancer was increasing and that the most common cause of thyroid disorders worldwide is iodine deficiency, leading to goitre formation and hypothyroidism

With respect to patients enrolled in the Medicare program, 111 of 719 people (15 percent) living in New Mexico had high serum TSH concentrations [7]

From the study of Jibril *et al* [8] on the prevalence of gestational thyroid disorders in Zaria, North-western, Nigeria. The objective of the work was to assess the prevalence of thyroid disorders among pregnant women in Zaria. The study showed that thyroid disorders among pregnant women in Zaria was high.

The prevalence of hyperthyroidism in women is between 0.5 and 2%, and is 10 times more common in women than in men in iodine-replete communities. The prevalence data in elderly persons show a wide range between 0.4 and 2.0% [9] and a higher prevalence is seen in iodine-deficient areas.

[10] looked at the thyroid disorders, etiology and prevalence. It was observed that major disorders (problems) of thyroid gland are hyperthyroidism and hypothyroidism, which have been reported in over 110 countries of the world with 1.6 billion people at risk and need some form of iodine supplementation. The study was limited to non- application of computer programs

[11] researched on ultrasound criteria for risk stratification of thyroid nodules in the previously iodine deficient area of Austria.

### B.Related work on thyroid disease using data mining algorithms

Mohmad [12] worked on suite of decision tree algorithms on cancer gene expression. The objective of the work was to compare the tree-based classification algorithms. Attribute selection techniques such as chi-square and Gain Ratio were used on ensembles learning methods: bagging, boosting, and Random Forest; enhanced classification accuracy of single decision tree due to the natures of its mechanism which generates several classifiers from one dataset and vote for their classification decision.

Ebru [13] used data mining models to classify thyroid disease. Decision tree algorithms to classify types of thyroid disease were compared and their performances according to performance metrics were analysed.

The study of prediction of thyroid disease using data mining techniques was carried out by Irina and Irina (2016) [4]. The authors analyzed and compared four classification models: Naive Bayes, Decision Tree, Multilayer Perceptron and Radial Basis Function

Network. The results indicated a significant accuracy for all the classification models. The best classification algorithm was Decision Tree model.

Gulmohamed *et al*[14] conducted a study on thyroid disease using data mining algorithms in which two main methods of data mining were observed such as clustering and classification. According to the perspective of Machine learning clustering method is unsupervised learning and tries to group sets of objects having relationship between them, whereas classification method is supervised and assigning objects to sets of predefined classes. The proposed system that the study entails was classification and clustering of thyroid disease.

## 3.0 METHODOLOGY

The architecture of the proposed predictive models of hormonal thyroid disease is presented in Figure 1. The major components of the architecture comprises of model construction and model evaluation. The attributes of the thyroid hormonal disease are passed to the classification algorithms which include REPTree, Random Forest and Decision Tree (C4.5) for the construction of the model. The Model Evaluation is another component in the model development process of thyroid hormonal disease which helps to find the best model that represents data and how well the chosen models work in the future.

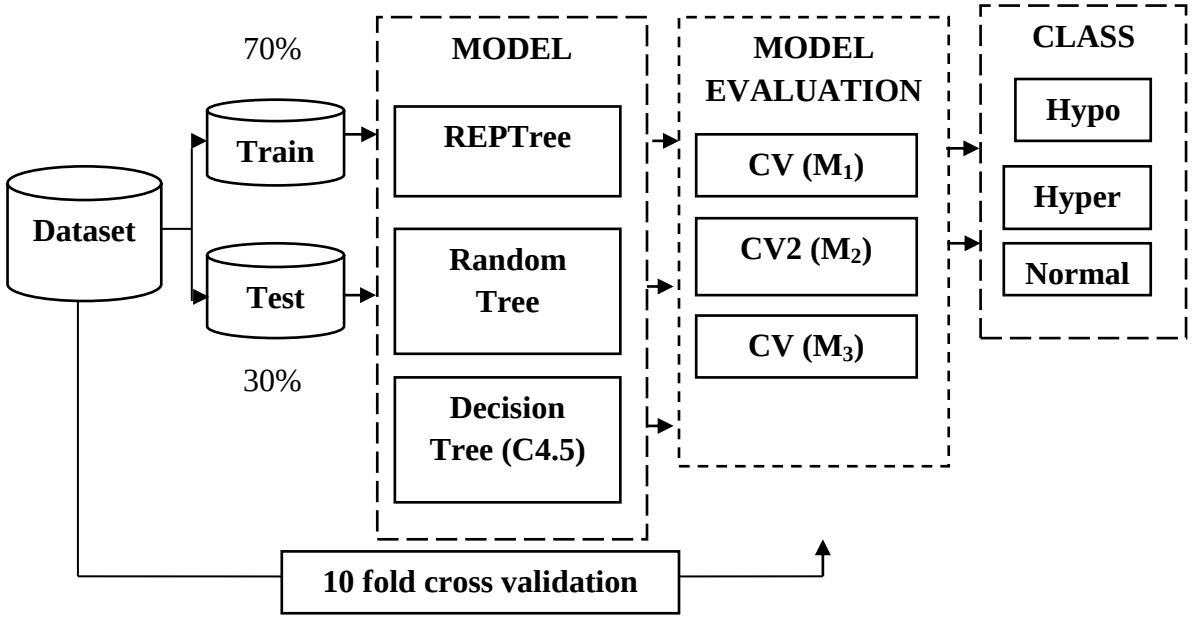


**Figure 1: Architecture of the thyroid hormonal disease**

### I. Thyroid dataset

The thyroid disease dataset was sourced from UCI Machine Learning Repository.
Thyroid disease dataset consists of the attributes and class as presented in Table 1

**Table 1: Data Set Attributes**

| s/n | Feature/predictors | Type |
|---|---|---|
| 1 | T3 Resin | Numeric |
| 2 | Serum Thyroxine | Numeric |
| 3 | Serum Triodothyronine | Numeric |
| 4 | Thyroid Stimulating Hormone | Numeric |
| 5 | Maximum Absolute Difference | Numeric |
| | Class | Normal (euthyroidism), hyper (hyperthyroidism) and hypo (hypothyroidism) |

## II.Features and class of Thyroid hormonal disease dataset

**T3 Resin:** This is a biochemical test, correctly called T3 Resin uptake test to assess thyroid – related protein in the blood. It shows the level of thyroid hormones available to the body in free form. Only the free form is utilizable by the body cells.

**Serum thyroxine:** This is the amount of thyroxine (T4) available in the blood. Both the total and free forms can be assayed. However, a free thyroxine test is considered more accurate than a total thyroxine test for checking thyroid function.

**Serum triiodothyronine:** This is the amount of triiodothyronine (T3) available in the blood. Its free form is essential for the proper metabolism of the organs in the body like the heart, muscle, brain, bone and digestive functions.

**Thyroid stimulating hormones (TSH):** This is a biochemical substances produced by a portion in the brain called the pituitary gland. It is carried in the blood to the thyroid gland which upon stimulation regulates the production of thyroxine (T4) and triiodothyroxine (T3) in a negative feedback format. Normal level of TSH will produce normal levels of T3 and T4. Higher level of TSH will produce lower levels of T3 and T4 (hypothyroidism) while lower level of TSH will produce higher levels of T3 and T4 (hyperthyroidism)

**Maximum absolute difference**

This is the absolute difference existing between the minimum and the maximum element from the array.

**Hypothyroidism:** This is a clinical condition in which the thyroid gland doesn't produce sufficient thyroid hormone.

**Hyperthyroidism:** This is a clinical condition in which the thyroid gland produces excessive thyroid hormones. All metabolic activities of the cells in the body become hyperfunctional.

## III.Training and test data

The dataset consists of 215 records and 5 features. The dataset was divided into: training data and testing data. 70% of 215 thyroid disease dataset was used for training data while the remaining 30% was used for testing data. The training data or cases were assumed to be represented as a pair [$x_1$, $x_2$, $x_3$, …$x_n$ → y] where $x_1$, $x_2$, $x_3$ …$x_n$ are vectors of attribute values describing some cases while y is the appropriate class or target.

## MODEL CONSTRUCTION- TREE-BASED ALGORITHMS

### 1. REPTree (Reduced – error pruning Tree) Classifier

RepTree is a fast decision tree learning algorithm. It builds a decision/regression tree by using information gain/variance and prunes it using reduced-error pruning (with backfitting) [15]. The algorithm uses decision tree technique which splits into branches starting from the root to the leaves. The algorithm prunes the tree using reduced-error with back fitting. The backfitting algorithm is an iterative procedure which is used to fit a generalized additive model. Backfitting algorithm is equivalent to the Gauss-Siedel method algorithm [16]

RepTree only sorts values for numeric attributes once. The missing values are replaced or fixed/death with by splitting the corresponding instances into pieces as in C4.5. RepTree has many options in its application such as maximum tree depth; minimum total weight of the instances in a leaf; minimum proportion of the variance on all the data that needs to be present at a node in order for splitting to be performed in regression trees, if pruning is performed; determining the amount of data used for pruning (One fold is used for pruning, the rest for growing the rules) and the seed used for randomizing the data.

### 2. Decision Tree (C4.5) classifier

A decision tree is used as a classifier for determining an appropriate action for a given case [17]. It determines the type of thyroid disease a patient could have whether it is hyper, hypo or normal. Information about a patient is given as vectors of attributes or input variables which include T3 Resin, Serum Thyroxin, Serum Triodothyronine, Thyroid Stimulating Hormone, Maximum Absolute Difference. The allowed actions are viewed as classes, which are in this case is hypo, hyper and normal. To find the appropriate class for a given patient (a person), it starts with the test at the root of the tree and keep following the branches as determined by the values of the features of the case at hand, until a leaf is reached.

The entropy of the class and each subset of the attribute/feature are computed using equations

$$E = -\sum_{i=1}^{n} p_i \log_2 p_i \quad (1)$$

Pi is the proportion of examples in thyroid that belong to the i-th class

n is number of classes.

Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the information gain that is presented in equation 2

Gain = Info (class) – Info (Attribute) (2)

Iterative Dichotomiser (ID3) uses Entropy and Information Gain (I.G.) to construct a decision tree. C4.5 is a successor of ID3 that uses Gain Ratio (G.R) which is computed using Equation 3

$$\text{Gain Ratio}_{\text{attribute}} = \frac{\text{Gain}_{\text{attribute}}}{\text{SplitInfo}_{\text{attribute}}} \quad (3)$$

### 3.Random Forest Classifier

Random forest is an ensemble learning method for classification or regression. Random forest is a type of supervised machine learning algorithm. It has the properties of ranking the feature or variable in a classification task. Random forest operates by

constructing a multitude of decision tree at training time and producing the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees [18]. In other way, it is an average of the predictions of each individual decision tree; in classification, it is the average of the most frequent prediction. The algorithm takes the average of many decision trees to arrive at a final prediction.

## MODEL EVALUATION

Cross validation is a method that evaluates model performance. The goal of cross validation is to have a good balance between the size and representation of data in the dataset both train set and test set.

Cross validation helps in finding the best model that represents data and how well the chosen model will work in the future. *n* fold cross-validation was used where the data are divided into *m* subsets of equal size. Models are built in m times, each time leaving out one of the subsets from training and use it as the test set.

## 4.0 RESULTS, DISCUSSION AND EVALUATION

The results of the three models using Decision Tree (C4.5), RepTree and Random tree on the split in proportion of 70% of training data and 30% of testing data are presented:

**Table 2: Performance metrics for the three classifiers**

| Metrics | REPTree | Random Forest | Decision Tree (C4.5) |
|---|---|---|---|
| Accuracy | 87.5 % | 89.0625 % | 87.5% |
| Precision | 0.876 | 0.898 | 0.879 |
| Recall | 0.875 | 0.891 | 0.840 |
| F measure | 0.871 | 0.886 | 0.858 |
| Kappa statistic | 0.7505 | 0.7749 | 0.7541 |
| Mean absolute error | 0.0984 | 0.0811 | 0.0911 |
| Root mean squared error | 0.2839 | 0.2062 | 0.2831 |
| Relative absolute error | 30.4191 % | 25.0978 % | 28.1746 % |
| Root relative squared error | 68.605 % | 49.815 % | 68.3941 % |

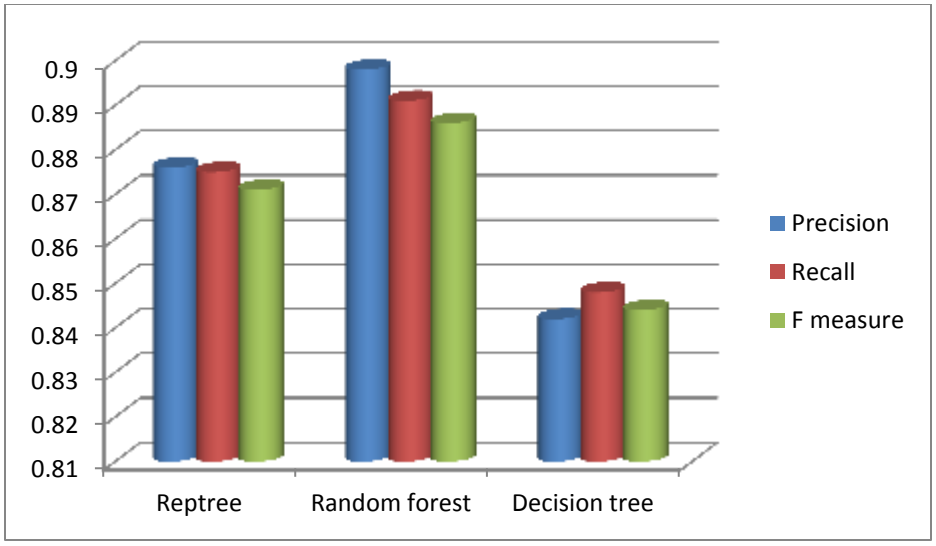


**Figure 2: Recall, precision and F measure of the Tree-based algorithms**

**Table 3: Result of the cross validation on the three tree-based classifier**

| Performance metric | REPTree | Random Forest | Decision Tree (C4.5) |
|---|---|---|---|
| Accuracy | 90.2326 % | 95.3488% | 91.16 % |
| Average Precision | 0.902 | 0.954 | 0.879 |
| Average Recall | 0.902 | 0.953 | 0.886 |
| F1 Score | 0.900 | 0.953 | 0.882 |
| Kappa statistic | 0.7802 | 0.8983 | 0.8119 |
| Mean absolute error | 0.0864 | 0.0517 | 0.0664 |
| Root mean squared error | 0.2506 | 0.1493 | 0.2268 |
| Relative absolute error | 27.5403 % | 16.473 % | 21.1632 % |
| Root relative squared error | 63.4941 % | 37.8281 % | 57.4634 % |

## 5.0 DISCUSSION ON THE PERFORMANCE METRICS OF THYROID HORMONAL DISEASE

The results of the predictive models of thyroid using learning algorithms such as: RepTree, Random Forest and Decision Tree (C4.5) are discussed:

When whole dataset is split into training data and test data in the proportion of 70% of the dataset for training and 30% of the dataset were for test data. The results showed that the accuracy of RepTree was 87.5%, Random Forest had the accuracy of 89.06% and Decision Tree (C4.5) was 87.5% respectively. F1 scores for RepTree, Random Forest and Decision Tree (C4.5) were 0.871, 0.886 and 0.858 respectively. This showed that Random Forest had highest accuracy and highest F1 score compared to RepTree and Decision Tree (C4.5). It further showed that Random Forest had the highest precision of 0.898 and recall of 0.891. The kappa statistic indicated 0.7749 for Random Forest which means that there was a strong agreement between the expected data and observed data. This shows that out of the three Tree-based models, Random Forest can be used in predicting the class category (hyperthyroidism, hypothyroidism or euthyroidism (normal)) of a patient who has hormonal thyroid disorder.

For the evaluation of the models, cross validation was induced on the dataset, in which the whole dataset (215 instances) were divided into 10 equal parts. One part was used as validation and other parts were used for training. The results showed that the accuracy of RepTree was 90.23%, Random Forest had the accuracy of 95.35% and Decision Tree (C4.5) was 91.16% respectively. F1 scores for RepTree, Random Forest and Decision Tree (C4.5) were 0.900, 0.0.953 and 0.882 respectively. This showed that Random Forest had the highest accuracy and highest F1 score compared to RepTree and Decision Tree (C4.5). It further showed that Random Forest had the highest precision of 0.954 and recall of 0.953 respectively. The kappa statistic for Random Forest was 0.8983 which indicated that there was a strong agreement between the expected data and observed data.

The best predictive model for thyroid hormonal dataset is Random Forest which yields F1 score of 0.8983 with accuracy of 95.35%. This is demonstrated below in a compact representation when comparing predictive models such as RepTree, Random Forest and Decision tree:

$\text{Model}_{GOOD} = M_{\text{random forest}} > M_{\text{decision tree (C4.5)}} > M_{\text{RepTree}}$

## 6.0 CONCLUSION

Many data mining algorithms are used in predicting the category of thyroid hormonal diseases in human body but one might be better than other. One effective way of analyzing this data is through data mining. The goal was to compare three tree-based algorithms and choose the best algorithm suited for predicting thyroid hormonal disease. The historical data on the patients show the rate of the chemical produced in the human body which predict if the chemical in the human body is normal, high or low.

The classification accuracy achieved by each method was compared to understand the effective method for thyroid hormonal disease.

## REFERENCES


1. *Taylor P. N., Albrecht D., Scholz A., Gutierrez-Buey G., Lazarus J. H., Dayan C. M. and Okosieme O. E. (2011). Global epidemiology of hyperthyroidism and hypothyroidism. British Medical Bulletin 2011; 99:39-51. doi: 10.1093/bmb/ldr030*
2. *Corbett E. D. (2018). The Real-World Benefits of Machine Learning in Healthcare. http:www/ www.healthcatalyst.com/ clinical-applications-of-machine-learning-in-healthcare*
3. *Ogbera A. O., Fasanmade O. and Adediran O. (2015). Pattern of thyroid disorders in the southwestern region of Nigeria.*
4. *Irina IoniŃă and Liviu IoniŃă (2016). Prediction of Thyroid Disease Using Data Mining Techniques. BRAIN. Broad Research in Artificial Intelligence and Neuroscience Volume 7, Issue 3, August 2016, ISSN 2067-3957 (online), ISSN 2068 - 0473*
5. *Bemben, D. A., Winn P, Hamm R. M, Morgan L, Davis A, and Barton E. (1994). Thyroid disease in the elderly. Part 1. Prevalence of undiagnosed hypothyroidism. J Fam Pract 38(6):577–582.*

6. *Vanderpump M P. J., Braverman L. E., Utiger R. D.(2005). The epidemiology of thyroid diseases. Werner and Ingbar's The Thyroid: A Fundamental and Clinical Text, edn. Philadelphia JB Lippincott-Raven. pg. 398-496*
7. *Lindeman RD, Schade DS, LaRue A, Romero LJ, Liang HC, Baumgartner RN, Koehler KM, Garry P. J. (1999). Subclinical hypothyroidism in a biethnic, urban community. J Am Geriatr Soc 47(6):703–709*
8. *Jibril Mohammed El-Bashir, Fayeofori Mpakabaori Abbiyesuku, Ibrahim Sambo Aliyu, Abdullahi Jibril Randawa, Rabiu Adamu , Sani Adamu, Shehu Abubakar Akuyam, Mohammed Manu, Hafsat Maiwada Suleiman, Rasheed Yusuf, Amina Mohammed (2016). Prevalence of gestational thyroid disorders in Zaria, north-western Nigeria. Issue 2 volume 9, Page : 51-55*
9. *GusseklooJ, Van Exel E, de Craen AJM (2004) . Thyroid status, disability and cognitive function, and survival in old age, JAMA, 2004, vol. 292, pg. 2591-99*
10. *Alam, Khan, M. Muzaffar Ali Khan and Shamim Akhtar (2002). Thyroid Disorders, Etiology and Prevalence. Journal of Medical Sciences, 2: 89-94*
11. *Christina Tugendsam, Veronika Petz, Wolfgang Buchinger, Brigitta Schmoll-Hauer, Iris Pia Schenk, Karin Rudolph, Michael Krebs and Georg Zettinig (Ultrasound criteria for risk stratification of thyroid nodules in the previously iodine deficient area of Austria - a single centre, retrospective analysis. BMC part of Springer nature https://doi.org/10.1186/s13044-018-0047-8*
12. *Mohmad Badr AlSnousy, Hesham Mohamed El-Deeb, Khaled Badran, Ibrahim Ali Alkhlil (2011). Suite of decision tree-based classification algorithms on cancer gene expression data. Egyptian Informatics Journal. Volume 12, issue 2. Pages 73-82*
13. *Ebru Turanoglu-Bekar, Gozde Ulutagay1 and Suzan Kantarcı-Savas (2016). Classification of Thyroid Disease by Using Data Mining Models: A Comparison of Decision Tree Algorithms. The Oxford Journal of Intelligent Decision and Data Science. Volume 2016, No. 2 (2016), Pages 13-28.*
14. *Gulmohamed Rasitha Banu, M. Baviya and Murtaza Ali (2015). A study on thyroid disease using data mining algorithm. International Journal of Technical Research and Applications 3(4):2320-8163*
15. *Mark Hall (2008). RepTree*
16. *Hastie, T. J. & Tibshirani, R. J. (1990). Generalized Additive Models. Monographs on Statistics and Applied Probability.43.*
17. *Jiawei Han, Micheline Kamber and Jian Pei (2011). Data mining: concepts and techniques (3rd edition)*
18. *Ho T. K (1998). The random subspace method for constructing decision forests (PDF). IEEE Transactions on Pattern Analysis and Machine Intelligence. 20 (8): 832–844.*