OPTIMAL FEATURE-BASED ENSEMBLE APPROACH FOR THYROID DISEASE CLASSIFICATION

[1]**C.SAJEE**, *M.E, Dept of Computer Science and Engineering*
[2]**M.MARKCO**, *M.E, (Ph.D), Assistant Prof, Dept of Computer Science and Engineering*
[12]*E.G.S. PILLAY ENGINEERING COLLEGE (AUTONOMOUS), NAGAPATTINAM.*

Abstract-Classification based Data mining plays important role in various healthcare services. In healthcare field, the important and challenging task is to diagnose health conditions and proper treatment of disease at the early stage. There are various diseases that can be diagnosed early and can be treated at the early stage. As for example, thyroid diseases the traditional ways of diagnosing thyroid diseases depend on clinical examination and many blood tests. The Main task is to detect disease diagnosis at the early stages with higher accuracy. Data mining techniques plays an important role in healthcare field for making decision, disease diagnosis and providing better treatment for the patients at low cost. Thyroid disease Classification is an important task. The purpose of this study is predication of thyroid disease using different **ensemble classification and feature selection techniques** and also to find the TSH, T3, T4 and more importance and feature selections in the dataset.

Keywords: Thyroid disease. Feature Selection. Ensemble Classification

I. INTRODUCTION

The thyroid is a little gland in the neck that produces thyroid hormones. It may produce too much or too small of these hormones. Hypothyroidism is a situation in which thyroid gland is not able to produce sufficient thyroid hormones. These hormones regulate metabolism of the body and further affects how the body uses energy. Lacking the accurate amount of thyroid hormones, body's normal functions start to slow down and body faces changes each day (hello, mood swings, happy, sad fatigue, depression, constipation, feeling cold, weight gain, muscle weakness, dry, thinning hair, slowed heart rate).

The hormones, total serum thyroxin (T4) and total serum tri iodothyronine (T3) are the two active thyroid hormones produced by the thyroid gland to control the metabolism of body. For the functioning of each cell and each tissue and organ in a right way, in overall energy yield and regulation and to generate proteins in the ordnance of body temperature, these hormones are necessary.

A. Thyroid Hormones

The thyroid gland produces are tri-iodothyronine (T3) and L-thyroxin (T4) and [6].The thyroid hormones regulates various metabolic activities such as generation of heat, the consumption of carbohydrates, protein and fats. The pituitary gland controls production of tri-iodothyronine and L-thyroxin hormones. The Thyrotropin-Stimulating Hormone [10] from pituitary gland is released when thyroid hormone is required and circulates through the bloodstream to reach thyroid gland. TSH then stimulates the thyroid glands for the production of T4 and T3 hormones [6].The production of thyroid hormone are controlled by the feedback system [6] of pituitary gland. The TSH production is less when T3, T4 are more in the circulation and TSH production is more when T3, T4 are less.

1) Hyperthyroidism

Increase production in the thyroid hormones causes hyperthyroidism. Graves' disease is one of the auto immune disorder that causes hyperthyroidism[9] .The symptoms[14] are dry skin , increase sensitivity to temperature, thinning of hair, weight loss, increase heart rate, high blood pressure, excess sweating, neck enlargement, nervousness, menstrual periods shorten, frequent gut movements and hands trembling [3].

2) Hypothyroidism

Decrease production in the thyroid hormones causes Hypothyroidism. The medical term hypo means deficient or less. The causes for hypothyroidism are inflammation and thyroid gland damage. The Symptoms includes obesity, low heart rate, and increase in cold sensitiveness, neck swelling, dry skin, hands numbness, hair problem, heavy menstrual periods and digestive problems. And these Symptoms may worsen over period if not treated [3].

B. Techniques

A. DECISION TREE

A decision tree contains 3 nodes i.e. root node, internal node and leaf node. The internal node performs test on given attribute, based on the test the classes are assigned to leaf nodes. The root node stays on the top of the decision tree. Decision trees have the ability to handle high dimensional data easily [28].

B. BOOSTING

Boosting is one of the Meta learning algorithms which focus on reducing bias. It has the capability of turning weak learners into strong ones. In boosting, the resulting models built, depend on the performance of past built models. During the process of boosting, the machine learning algorithm looks for to find misclassified instances, applies extra weights on each of the misclassified instances and then builds the fresh training data set for new model.

C. BAGGING

Bagging is used in statistical classification and regression that improves the stability and the accuracy of deployed machine learning algorithm. Bagging is very useful in avoiding over fitting and reduces variance. Bagging uses the model averaging approach for predicting the results.

D. K-Nearest Neighbour (K-NN)

A k-nearest neighbour often abbreviated as k-NN algorithm. It is the data classification that estimates likely as a data point into the member of one group or into the other depending on grouping the data points that may be nearest. Thek-nearest-neighbour is also called as a "lazy learner "algorithm that not be built on a model that is using in a training set until the query of the data set is performed[9].

Algorithm: The algorithm is in the case, is classified by the majority of vote to its neighbours, and with the case being assigned to the class, the most common among its K nearest neighbours. Measured by a distance function. If the value K = 1, then the Case value is simply assigned to the class of its nearest Neighbour. The three distance measures are noted as valid continuous Variables.

$$(x + Sa)^n = \sqrt{\sum_{i=i}^{k} (xi - yi)2}$$

E. Naive Bayes

It is the simple classification algorithm for predicting modelling with clear semantics, representing and the probabilistic learning method based on Bayesian theorem. Naive classifier assumes the value of the one attribute is not dependent on the value of another attribute and it assumes that the presence or absence of particular attribute of the prediction process does not affect. Suppose there are m classes say K1,K2….Kn having a unidentified data sample X, Naïve Bayesian classifier will predict an unknown sample X to the class Ki on the basis of the classes having highest probability[3] [9].

P (Ki |X > P (Kj |X) for 1≤j≤m, j ≠

F. Support Vector Machine

Support Vector Machine is one of the managed machine learning algorithm used for both the classification and regression issues and it is usually used for a bit of arrangement problems. The estimation of selected organize is of the each half being the estimation. Then the tendency to perform characterization by finding the hype-plane is completely having categories. The positive value represents terrorist and also as the model says it's a non-terrorist. Idea about the costs that having a mis-classified actual positive value is very high there.

II. LITERATURE SURVEY

Senthil Kumar Mohan et al, [1] proposed Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques in which strategy that objective is to finding critical includes by applying Machine Learning bringing about improving the exactness in the expectation of cardiovascular malady. The expectation model is created with various blends of highlights and a few known arrangement strategies. We produce an improved exhibition level with a precision level of 88.7% through the prediction model for heart disease with hybrid random forest with a linear model (HRFLM) they likewise educated about Diverse data mining approaches and expectation techniques, Such as, KNN, LR, SVM, NN, and Vote have been fairly famous of late to distinguish and predict heart disease.

SonamNikhar et al [2] has built up the paper titled as Prediction of Heart Disease Using Machine Learning Algorithms by This exploration plans to give a point by point portrayal of Naive Bayes and decision tree classifier that are applied in our examination especially in the prediction of Heart Disease. Some analysis has been led to think about the execution of prescient data mining strategy on the equivalent dataset, and the result uncovers that Decision Tree beats over Bayesian classification system.

AditiGavhane, GouthamiKokkula, IshaPandya, Prof. Kailas Devadkar (PhD),[3] Prediction of Heart Disease Using Machine Learning, In this paper proposed system they used the neural network algorithm multi-layer perception (MLP) to train and test the dataset. In this algorithm there will be multiple layers like one for input, second for output and one or more layers are hidden layers between these two input and output layers. Each node in input layer is connected to output nodes through these hidden layers. This connection is assigned with some weights.

Abhay Kishore et al, [4] developed Heart Attack Prediction Using Deep Learning in which this paper proposes a heart attack prediction system using Deep learning procedures, explicitly Recurrent Neural System to predict the probable prospects of heart related infections of the patient. Recurrent Neural Network is a very ground-breaking characterization calculation that utilizes Deep Learning approach in Artificial Neural Network. The paper talks about in detail the significant modules of the framework alongside the related hypothesis .The proposed model deep learning and data mining to give the precise outcomes least blunders. This paper gives a bearing and point of reference for the advancement of another type of heart attack prediction platform .Prediction stage.

LakshmanaRao et al, [5] Machine Learning Techniques for Heart Disease Prediction in which the contributing elements for heart disease are more (circulatory strain, diabetes, current smoker, high cholesterol, etc...). So, it is difficult to distinguish heart disease. Different systems in data mining and neural systems have been utilized to discover the seriousness of heart disease among people. The idea of CHD ailment is bewildering, in addition, in this manner; the disease must be dealt with warily. Not doing early identification, may impact the heart or cause sudden passing.

III. PROPOSED APPROACH

In healthcare services data mining technique is mainly used for making decision, disease diagnosing and giving better treatment to the patients at comparately low cost. Classification of thyroid disease plays is an important task in the prediction of disease. Dimensionality reduction may be done as a future work so that number of blood test the thyroid will be reduced and also time required diagnosing disease. The thyroid Dataset is taken from UCI data repository site. The Database consists of thyroid patient records. The Patients record is having different attributes described in the data set description and different data mining techniques are applied to get the predication of thyroid disease. Data mining Algorithms such as ensemble models bagging, boosting, feature selection and sampling techniques are considered for the study.

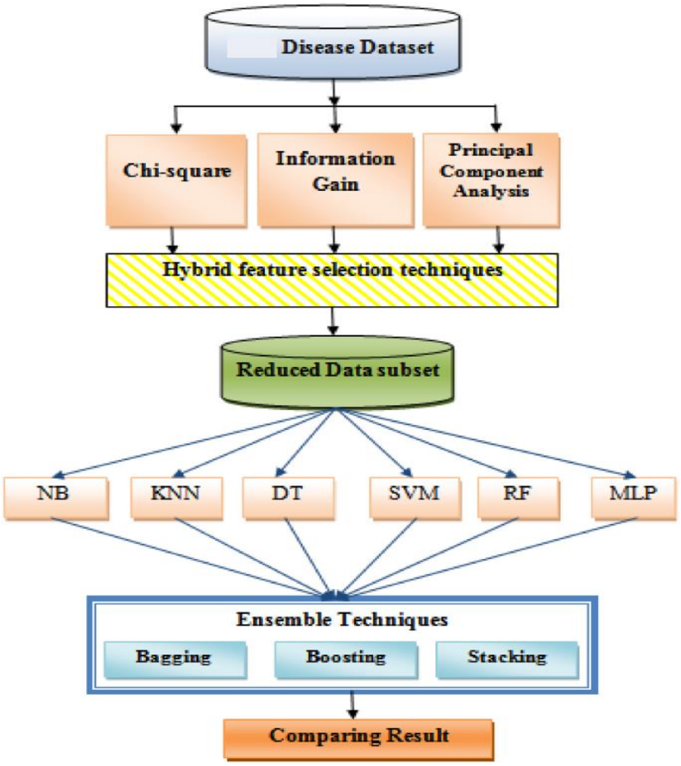


Fig.1 System Architecture

IV. CLASSIFICATION

It is a process of categorising data into given classes. Its primary goal is to identify the class of our new data.

4.1 Machine learning algorithms for classification

Research on data mining has led to the formulation of several data mining algorithms. These algorithms can be directly used on a dataset for creating some models or to draw vital conclusions and inferences from that dataset. Some popular data mining algorithms are Decision tree, Naïve Bayes, k-means, artificial neural network etc.

4.2 Decision Tree: Decision Tree Analysis is a general, predictive modelling tool that has applications spanning a number of different areas. In general, decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

4.3 Naive Bayes (NB): It is a simple technique for constructing classifiers. It is a probabilistic classifier based on Bayes' theorem. All Naive Bayes classifiers assume that the value of any particular feature is independent of the value of any other feature, given the class variable. Bayes theorem is given as follows: **P (C|X) = P (X|C) * P(C)/P(X)**, where X is the data tuple and C is the class such that P(X) is constant for all classes.

4.4 Random Forest: Random Forests are an ensemble learning method (also thought of as a form of nearest neighbour predictor) for classification and regression techniques. It builds multiple decision trees and then merges them together in-order to get more accurate and stable predictions. It constructs a number of Decision trees at training time and outputs the class that is the mode of the classes output by individual trees.

4.5 KNN: KNN algorithm is one of the simplest classification algorithms and it is one of the most used learning algorithms. KNN is a non-parametric, lazy learning algorithm. Its purpose is to use a dataset in which the data points are separated into several classes to predict the classification of a new sample point. A KNN algorithm uses a data and classifies new data points based on a similarity measures (e.g. distance function, error rate). Classification is done by a majority vote to its neighbours. The data is assigned to the class which has the most nearest neighbours. As we increase the number of nearest neighbours, the value of k, accuracy may increase.

4.6 Logistic Regression: Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.Logistic Regression is much similar to the Linear Regression except that how they are used.

SMOTE Technique: SMOTE (synthetic minority oversampling technique) is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them. SMOTE synthesizes new minority instances between existing minority instances. It generates the virtual training records by linear interpolation for the minority class. These synthetic training records are generated by randomly selecting one or more of the k-nearest neighbors for each example in the minority class. After the oversampling process, the data is reconstructed and several classification models can be applied for the processed data.

Data Pre-processing

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Pre-processing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible.

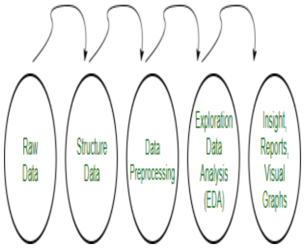

Raw Data
Structure Data
Data Preprocessing
Exploration Data Analysis (EDA)
Insight, Reports, Visual Graphs

Fig.2 Data Pre-processing

Need of Data Pre-processing

For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format. For example, Random Forest algorithm does not support null values; therefore to execute random forest algorithm null values have to be managed from the original raw data set. Another aspect is that data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in one data set and best out of them is chosen.

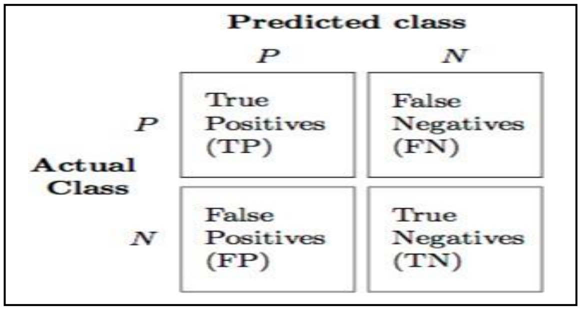
V. RESULTS AND DISCUSSIONS

This research work done in data mining tools in window environment. We have used various classification techniques like random Forest, Naive Bayes and K-NN for classification of thyroid disease. We have proposed new ensemble model that is combination of Random Forest, Naive bayes and KNN which gives better classification accuracy as 93.55% compare to other individuals models.

5.1 PEFORMANCE MEASURES

CONFUSION MATRIX

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.



		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

A true positive (tp) is a result where the model predicts the positive class correctly. Similarly, a true negative (tn) is an outcome where the model correctly predicts the negative class.

A false positive (fp) is an outcome where the model incorrectly predicts the positive class. And a false negative (fn) is an outcome where the model incorrectly predicts the negative class.

Accuracy: It is the ratio of the correctly classified packets (normal or attacks) to the total dataset. It can be calculated as:

Accuracy

$TP+TN$

$TP+TN+FP+FN$

Precision: It is the ratio of correctly classified attacks to the total number of identified attacks. It can be calculated as:

Precision

TP

$TP+FP$

Recall: It is the ratio of accurately classified attacks to the total number of attacks in the test dataset. It can be calculated as:

TP

$TP+FN$

F1-Score: It is the average of the precision and the Recall with a weight of 2. It can be calculated as:

Precision× Recall

2*----------------------------

Precision +Recall

False Positive Rate (FPR) is the number of normal connections that are recognized as an attack on the total number of normal connections.

FP

$TN+FP$

5.2 Evaluation

- The confusion matrix, also known as the error matrix, is used to evaluate the accuracy of the model.

Accuracy: It is the ratio of the correctly classified packets (normal or attacks) to the total dataset. It can be calculated as:

Accuracy

$TP+TN$

$TP+TN+FP+FN$

Precision: It is the ratio of correctly classified attacks to the total number of identified attacks. It can be calculated as:

Precision

TP

$TP+FP$

Recall: It is the ratio of accurately classified attacks to the total number of attacks in the test dataset. It can be calculated as:

TP

$TP+FN$

F1-Score: It is the average of the precision and the Recall with a weight of 2. It can be calculated as:

Precision× Recall

2*----------------------------

Precision +Recall

False Positive Rate (FPR) is the number of normal connections that are recognized as an attack on the total number of normal connections.

$$\frac{FP}{TN+FP}$$

Fig.3 Outputs-Algorithm Comparison

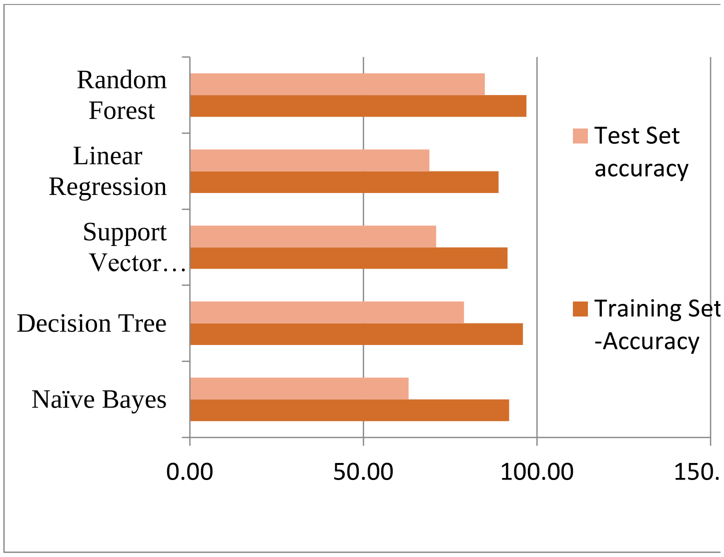
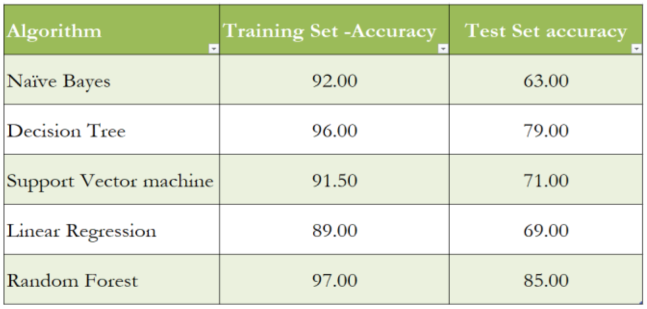


Table 1: Accuracy Table



Algorithm	Training Set -Accuracy	Test Set accuracy
Naïve Bayes	92.00	63.00
Decision Tree	96.00	79.00
Support Vector machine	91.50	71.00
Linear Regression	89.00	69.00
Random Forest	97.00	85.00

Fig.4 Test Set Accuracy-Pie Chart

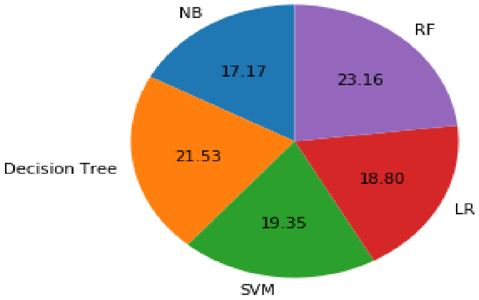


Fig.5 Bar plot-Accuracy

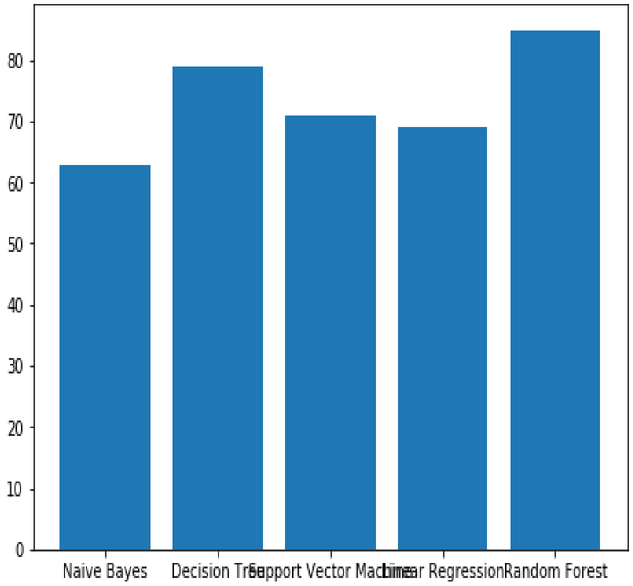


Table 2: Performance measures of proposed ensemble model

Accuracy	93.49%
Sensitivity	93.55%
specificity	91.78%

Table 3: Feature optimization technique on proposed ensemble model

Feature selection technique	Number of features	Name of features	Accuracy (%) after feature selection
Optimize Selection	3	TSH, T3 measured, T3	97.61

VI. CONCLUSION

Thus this survey is needful to identify how to predict the thyroid disorder at earlier stage using data mining techniques. Data mining classification algorithms are used to diagnose the thyroid problems and gives different level of accuracy for each technique. These techniques help to minimize the noisy data of the patient's data from the data bases. Data mining Algorithms such as KNN, Naive Bayes, Support vector machine, ensemble methods are considered and for the study. These various algorithm results are based on speed, accuracy and performance of the model and cost for the treatment. Also these classifications of effective data are helps to find the treatment to the thyroid patients with better cost and facilitate the management. The ensemble techniques are applied on the hypothyroid and hypothyroid and sick thyroid dataset and it is also to determine the positive and the negative values from the entire dataset. The experimental result provides, when compared to male and female dataset, females are more affected than male. The improved accuracy, precision and recall by comparing the Decision tree, Support vector Machine and ensemble methods. Further enhancement has been made by using the various optimization algorithms or rule extraction algorithms. The future work is applied on validating the multiple disease dataset simultaneously like heart disease, diabetics, etc.

REFERENCES

[1]. Ahmed, Jamil, and M. Abdul RehmanSoomrani.“TDTD: Thyroid disease type diagnostics.” 2016 International Conference on Intelligent Systems Engineering (ICISE). IEEE, 2016.

[2]. Ammulu, K., and T. Venugopal. “Thyroid data prediction using data classification algorithm.” Int.J. Innov. Res. Sci. Technol 4.2 (2017): 208-212.

[3]. Begum, Amina, and A. Parkavi. “Prediction of thyroid Disease Using Data Mining Techniques.”2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS) IEEE, 2019.

[4]. Chang, Chuan-Yu, Ming-Feng Tsai, and Shao-JerChen. “Classification of the thyroid nodules using support vector machines.”2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence).IEEE, 2008.

[5]. Dov, David, et al. “A Deep-Learning Algorithm for Thyroid Malignancy Prediction From Whole Slide Cytopathology Images.” arXiv preprintarXiv:1904.12739 (2019).

[6]. Geetha, K., and S. SanthoshBaboo. “An empirical model for thyroid disease classification using evolutionary multivariate Bayesian prediction method.”Global Journal of Computer Science and Technology (2016).

[7]. Ioniţa, Irina, and LiviuIoniţa. “Prediction of thyroid disease using data mining techniques.”BRAIN. Broad Research in Artificial Intelligence and Neuroscience 7.3 (2016): 115-124.

[8]. Kousarrizi, MR Nazari, F. Seiti, and M. Teshneh lab.“An experimental comparative study on thyroid disease diagnosis based on feature subset selection and classification.” International Journal of Electrical & Computer Sciences IJECS-IJENS12.01 (2012): 13-19.

[9]. Margret, J., B. Lakshmipathi, and S. Aswani Kumar.“Diagnosis of thyroid disorders using decision tree splitting rules.” International Journal of Computer Applications 44.8 (2012): 43-46.

[10]. Prasad, V., T. SrinivasaRao, and M. SurendraPrasad Babu. “Thyroid disease diagnosis via hybrid architecture composing rough data sets theory and machine learning algorithms.” Soft Computing20.3 (2016): 1179-1189.

[11]. Raisinghani, Sagar, et al. “Thyroid Prediction Using Machine Learning Techniques.” International Conference on Advances in Computing and Data Sciences. Springer, Singapore, 2019.

[12]. ShaikRazia, P.SwathiPrathyusha, N.VamsiKrishna, N. SathyaSumana.” A Comparative study of machine learning algorithms on thyroid disease prediction”, International Journal of Engineering &Technology, 7 (2.8) (2018) 315-319.

[13]. ShaikRazia, P.SwathiPrathyusha, N.VamsiKrishna, N.SathyaSumana, "A Comparative study of machine learning algorithms on thyroid disease prediction" International Journal of Engineering &Technology (UAE), vol 8, 7 (2.8) (2018) 315-319.

[14]. Visser, Theo J. "Regulation of Thyroid Function, Synthesis and Function of Thyroid Hormones."Thyroid Diseases: Pathogenesis, Diagnosis and Treatment (2018): 1-30.

[15]. Yadav, Dhyan Chandra, and Saurabh Pal. "ToGenerate an Ensemble Model for Women Thyroid Prediction Using Data Mining Techniques." Asian Pacific Journal of Cancer Prevention 20.4 (2019):1275-1281.

[16] RoshanBanu D, and K.C.Sharmili "A Study of Data Mining Techniques to Detect Thyroid Disease" International Journal of Innovative Research in Science, Engineering and Technology (Vol. 6, Special Issue 11, September 2017

[17]Irina IoniŃă and LiviuIoniŃă" Prediction of Thyroid Disease Using Data Mining Techniques" The Classification Technique for Talent Management using SVM, (ICCEET), 978-1- 4673-02l0-4/12, pp. 959- 963, 2017.

[18] KhushbooTaneja, ParveenSehgal, Prerana "Predictive Data Mining for Diagnosis of Thyroid Disease using Neural Network" International Journal of Research in Management, Science & Technology (E-ISSN: 2321- 3264) Vol. 3, No. 2, April 2016.

[19] HanungAdiNugroho, Noor AkhmadSetiawan, Md. DendiMaysanjaya," A Comparison of Classification Methods on Diagnosis of Thyroid Diseases" IEEE International Seminar on Intelligent Technology and Its Applications,2017

[20] M.N Das ,Brojo Kishore Mishra ,Shreela Dash," Implementation of an Optimized Classification Model for Prediction of Hypothyroid Disease Risks" 2017 IEEE Conference on Big Data and Analytics (ICBDA)