Prediction of Thyroid Disease Type Using Combination of Different Machine Learning Techniques

[1]Anuradha Shyam, [2]Mohanrao Mamdikar, [3]Pooja Patre
[1]M.Tech. scholar , [2]Assistant Professor, [3]Assistant Professor
[1]Department of CSE
[1]Vishwavidyalaya Engineering College, Lakhanpur,Chhattisgarh, India

Abstract—: **The thyroid gland is one among the most necessary organs in our body. It secretes thyroid hormones that are responsible for controlling metabolism. The less secretion hormone causes hypothyroidism and far secretion of thyroid causes hyperthyroidism. For deciding, data processing technique is especially utilized in health care sectors, sickness identification and giving better treatment to the patients. Here, in this paper we have used two different datasets thyroid disease from UCI and various machine learning techniques have been applied for prediction of thyroid disease. Finally results are compared on the basis of confusion matrix. After performing experiments we observed that the use of neural network models and decision tree classifier produces good results as both classifier produces accuracy more than 99%.**

Keywords:- **Decision Tree, KNN, Neural Network, Regression, Support Vector Machine, Thyroid Diseases.**

1. INTRODUCTION

Disease identification could be a terribly advanced and tedious task; because it needs many expertise and data. one in all the standard ways in which for identification is doctor's examination or variety of blood tests. the most task is to supply malady identification at early stages with higher accuracy. data processing plays a significant role in medical field for malady identification. consistent with an analysis whereas one in 10 adults in India's individuals is affected by hypothyroidism. This estimation is found on the premise of an analysis conduct by Indian thyroid society. The study additionally alert for thyroid and thyroid is ninth hierarchical compared to different kind common sickness like asthma attack, cholesterol, depression, diabetes etc. medical practitioner say that thyroid are same as different disorders, however, the investigation population area unit conscious of thyroid disorders, understand that there area unit diagnostic tests for locating of this disease[1,2]. The thyroid releases 2 principal hormones. the primary is termed thyroid hormone (T4) and therefore the different one is thyroid hormone (T3) into the blood stream. the most functions of the thyroid hormones area unit to manage the expansion rate of metabolism. There are 2 common issues of thyroid disorder: thyrotoxicosis and hypothyroidism. the primary one releases an excessive amount of hormone into the blood stream and therefore the second releases too low thyroid hormones to the blood stream[3][4].

Classification algorithms are important class of supervised machine learning algorithms. These algorithms need a really massive training sets. These training information sets are consisting of the many features or attributes that describe the individual sample. Since we have a tendency to do supervised learning algorithmic rule. All of the training set area unit tagged properly. The classification algorithms like call trees and support vector machines (SVM), develop model with these information with many alternative parameters. after we have a replacement untagged sample, we are able to use the model to predict the label of the new sample. These techniques are used for malady identification to assist doctor to effectively label the new case[5][6]. a call tree is one in all the terribly effective machine learning classifier wherever the algorithmic rule makes a tree structure, wherever each non-leaf node denotes a take a look at on an attribute, every branch performs associate outcome of the take a look at and every leaf node holds a category label. call tree is incredibly easy and effective classifier thus this algorithmic rule may be utilized in many application areas like medication, money analysis, physical science and biology for classification [7][8].

2. LITERATURE REVIEW

Qureshi M.A. et.al [12] proposes totally different|completely different} decision trees alorithms classify different thyroid-related diseases. 1st they applied feature reduction ways to remove ten inappropriate features from twenty nine features.They found the between ninety seven.43% and 99.18% for various thyroid tasks. Ahmed J. et.al. [13] planned a system TDTD that may be a unique technique) for prediction of missing values in medical datasets, finally for classification they used support vector machine. They achieved accuracy of ninety five.7% on UCI dataset of thyroid unwellness.Azar A.T. et.al. [14] planned a fuzzy based mostly technique for feature choice and have reduction, finally for classification purpose fuzzy agglomeration is employed they achieved accuracy between ninety eight to ninety nine.5% on Classification of information from the University of Golden State, Irvine (UCI) machine learning information set repository was performed to guage the effectiveness of the Neural-fuzzy classifier on real-world information, and to facilitate comparison with different classifiers[4]. The dataset contains three categories and 215 samples. These categories ar assigned to the values that correspond to the hyper-, hypo-, and traditional perform of the endocrine gland. All samples have 5 features. The results indicated that the classification accuracy while not feature choice was ninety eight.6047% and ninety seven.6744% throughout training and testing phases, severally with RMSE of 0.02335. once applying feature choice algorithmic rule, LHNFCSF achieved 100 percent for all cluster sizes throughout training part. However, within the testing part LHNFCSF achieved eighty eight.3721% mistreatment one cluster for every category, 90.6977% using 2 clusters, 91.8605% using 3 clusters and ninety seven.6744% mistreatment four clusters for every category and twelve fuzzy rules[14].

Rishita bestowed a research within which, LDA algorithmic rule is employed to predicate thyroid sickness. an information set with twenty nine options downloaded from UCI repository web site is employed for the experimental purpose [9][10], entire work is disbursed with maori hen open supply computer code below Windows 7environment. K-fold validation is additionally performed. There ar in total 3772 records within the hypo thyroid information set. All the records ar classified as negative, stipendiary hypothyroid, primary hypothyroid or secondary hypothyroid. In our experiment information is provided to classifier of LDA algorithmic rule. during this paper we've applied LDA data processing classification techniques is employed to classify the hypothyroid unwellness. K-fold cross validation is additionally performed. The LDA algorithmic rule offers ninety nine.62% accuracy with k=6 folds cross validation[16]. Ahemed et.al. planned a technique for thyroid sickness diagnosing mistreatment Hybrid call web supported ANFIS, k-NN and knowledge Gain technique. The novelty of this technique is that it's a hybrid system, comprised of feature choice method mistreatment data gain technique that decreases computation time and will increase the accuracy of the ensuing model, k-NN Imputation for missing information values and ANFIS system, that maximize the generalization capability of our thyroid diagnosing system. Performance comparison of our planned system to antecedently introduced ways like MLP, LVQ, RBF, PPFNN, MLP with back-prop, MLP quick back-prop, LDA, C4.5-1, C4.5-2, C4.5-3, MLP, DIMLP, AIRS AIRS with fuzzy weighted, ESTDD, MLNN with luminous flux unit, PNN, LVQ, GDA-WSVM, and FS-PSO-SVM[7, 9-13] were terribly low. however the results tried this planned diagnosing system has higher performance than non-hybrid schemes[17].

S umadevi et.al. planned a technique for thyroid sickness diagnosing mistreatment KNN. during this work, they need applied ANN, k-nearest neighbour and fuzzy classifiers on 2 normal thyroid datasets. The results show that the ANN classifier presented a really high classification accuracy of ninetieth mistreatment geometer and Manhattan Distances severally. The high accuracy encourages United States to validate the system employing a larger and a unique medical dataset so as to determine its clinical pertinency to help doctors in thyroid classification and future treatment regime[18]. Rastogi A., it's seen that feed forward neural network is with success used for the diagnosing of thyroid unwellness. Thyroid sickness identification is a very important nonetheless tough task from each clinical diagnosing and applied math classification purpose of read. The poor performance of the normal model based mostly applied math ways thanks to large number of reticular patient attributes in addition as extraordinarily unbalanced teams within the thyroid diagnosing drawback complicate the link between these attributes and therefore the patient true cluster membership. Artificial neural network may be a versatile modelling technique for advanced perform mapping, show promise within the thyroid sickness diagnosis[19].

3. METHODOLOGY

Figure 1 shows the overall methodology of our system where we are using 2 different datasets downloaded from UCI repository of machine learning. Then these datasets are proposed by removing null values, missing data etc. finally classifiers are applied by training and testing the system.

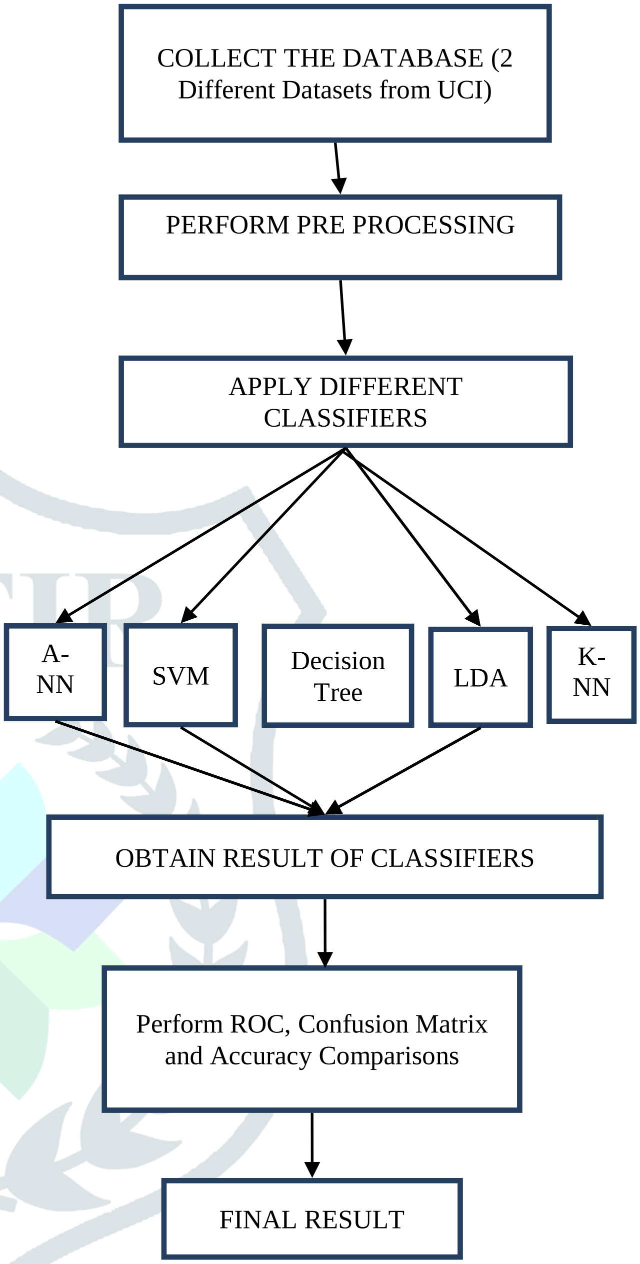

COLLECT THE DATABASE (2 Different Datasets from UCI)
PERFORM PRE PROCESSING
APPLY DIFFERENT CLASSIFIERS
A-NN
SVM
Decision Tree
LDA
K-NN
OBTAIN RESULT OF CLASSIFIERS
Perform ROC, Confusion Matrix and Accuracy Comparisons
FINAL RESULT

Figure 1: Flow chart of Proposed Methodology

A. Data collection

First Dataset:

Thyroid data set is collected from UCI repository[9] was performed to evaluate the effectiveness of the Neural-fuzzy classifier on real-world data, and to facilitate comparison with other classifiers. [4]. We are using 2 different datasets for experiments, The first dataset contains 3 classes and 215 samples. These classes are assigned to the values that correspond to the hyper-, hypo-, and normal function of the thyroid gland.

Second Dataset:

Second dataset The original thyroid disease (ANN-thyroid) dataset from UCI machine learning repository is a classification dataset, which is suited for training ANNs. 3428 instances. There are 21 features, 15 of them are categorical and 6 of them are real attributes. The problem is to determine whether a patient referred to the clinic is hypothyroid. Therefore three classes are built to decide whether a patient has thyroid over--, normal-- or under function.

.

B. Feature extraction

Features of First Dataset:

First data set content 5 major features that plays a vital role in detection of thyroid disease and these features are:

1. T3-resin uptake test (A percentage).
2. Total serum thyroxin as measured by the isotopic displacement method.
3. Total serum triiodothyronine as measured by radioimmunology assay.
4. Basal thyroid-stimulating hormone (TSH) as measured by radioimmunology assay.
5. Maximal absolute difference of TSH value after injection of 200 µg of thyrotropin- reeleasing hormone as compared to the basal value.

Features of Second dataset:

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-thyroidism, hypo thyroidism and normal function of the thyroid gland.

C. Classification of Disease:

Here I will use all the samples taken from UCI machine repository and the 5 features belong to such classes. Another dataset taken contains the 21 features for classification and more than 7000 instances. After then I will individually classify them by using ANN classifier, SVM classifier, Decision Tree and Linear Discriminate Analyser.

I. Algorithm applied Support Vector Machine:

```
Input: initialize subset S = { 1,2,3.....}
Output: Rank list according to smallest weight R
Step1: Initially defined R = {  }.
Step2: Repeat step 3 to 8 until G is not empty.
Step3: Train support vector machine model using G.
Step4:Compute weight W vector for SVM
Step5: Compute Rank R= W*W
Step6: Rank features and sort accordingly
          $Rank_{new}$ = Sort (Rank);
Step7: Update feature rank list
          Update R = R + G ($Rank_{new}$)
Step8: Eliminate feature with smallest rank
          Update G = G- G($Rank_{new}$)
Step9: End
```

II. Algorithm used for Naive Bayes classification:

The dataset is divided into two different categories like feature matrix and response vector.

Step1: The feature matrix consists each and every vector that is nothing but row of dataset in which every vector comprises of the estimate of dependent features.

Step2: Response vector contains the esteem of class variable (prediction or yield) for each vector which is nothing but a row of the feature matrix.

Bayes' Theorem finds the likelihood of an event happening given the likelihood of another event that has just happened. Bayes' hypothesis is expressed numerically as the following condition:

P(A/B) = P(B/A)P(A) / P(B) (1)

Step1: Convert the data set into a frequency table.
Step2: Create like hood table by finding the probabilities.
Step3: Now, use naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

III. Algorithm used for Decision Tree classification:

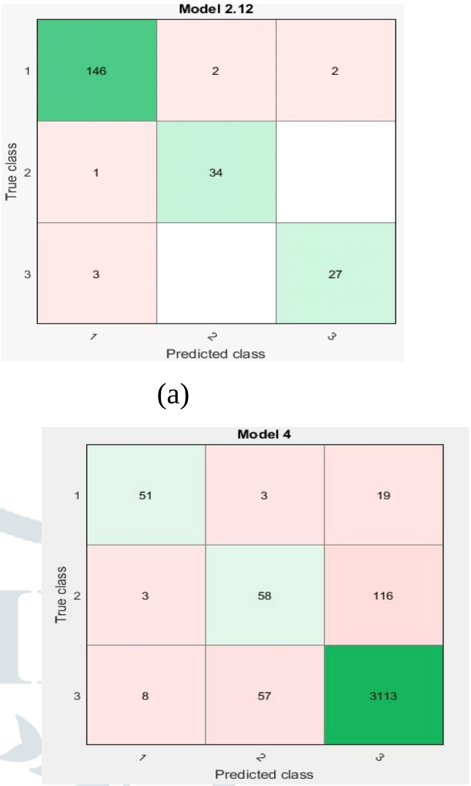
Step1: Place the best attribute of the dataset at the root of the tree
Step2: Split the training set into subsets.
Step3: Repeat step1 and step2 on each subset until you find the leaf nodes in all the branches of the tree

D. Performance Evaluation and Comparison

After classification I will evaluate the performance that means in simple words I can say after getting results from ANN, SVM, decision Tree, KNN classification I will compare result of all three i.e. evaluated from ANN, SVM , decision Tree, KNN classification. After comparison whose performance is highest among three will be my final result**.** A confusion matrix consist actual and predicted classification, computed by a classification scheme. The confusion matrix can also be viewed as contingency table or an error matrix, it is a specific table that depicts the performance of a classifier.

4. RESULT ANALYSIS

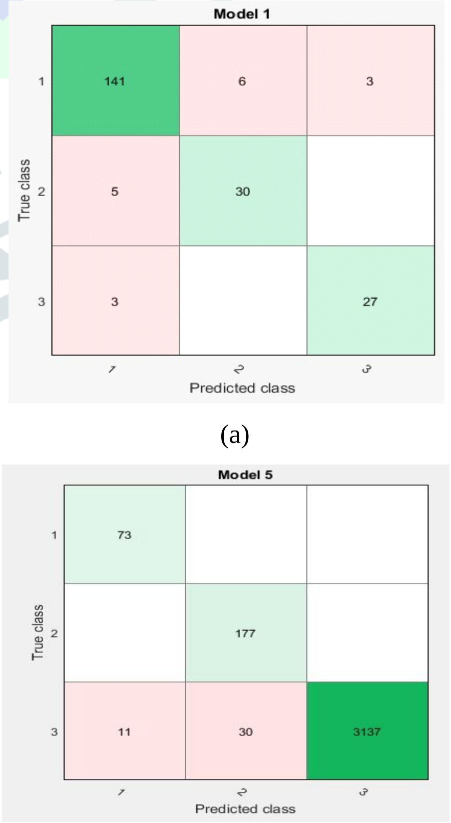
After performing lots of experiments we got confusion matrices shown in figures below. From these confusion matrices we have calculated the accuracies of different classifiers.



(a)

(b)

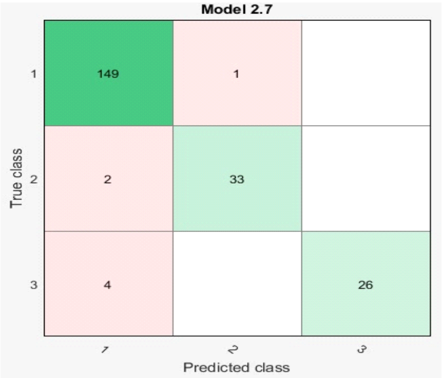
Figure 2: Confusion matrix of K- nearest neighbor on 5 features(a) and 21 features(b)



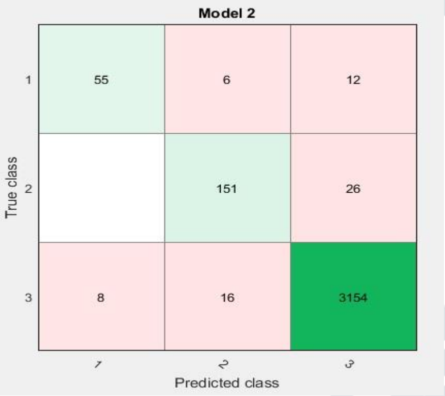
(a)

(b)

Figure 3: Confusion matrix of Decision Tree on 5 features(a) and 21 features(b)

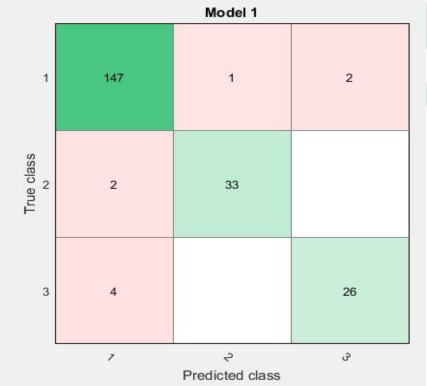


(a)

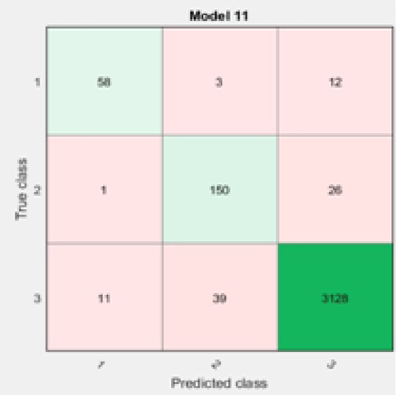


(b)

Figure 4: Confusion matrix of Support Vector Machine on 5 features (a) and 21 features(b)

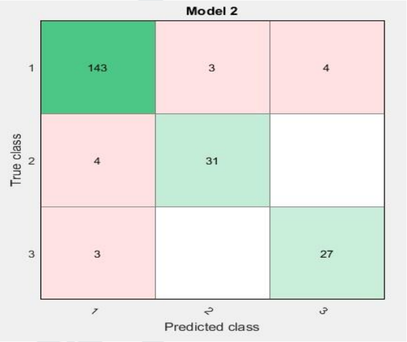


(a)

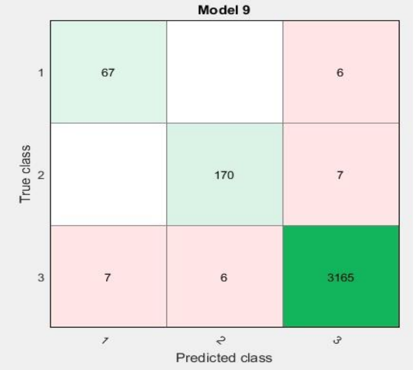


(b)

Figure 5: Confusion matrix of Cubic Support Vector Machine on 5 features (a) and 21 Features (b)



(a)



(b)

Figure 6: Confusion matrix of Medium Decision Tree on 5 features(a) and 21 Features(b)

Form table 1 we can observed that most of the classifiers performs well in different datasets as per our experiments we can see that highest accuracy obtained from decision tree classifier on dataset 2 i.e. 99.2 % while on dataset 1 highest accuracy obtained was 99.14% using neural networks.

Accuracies are almost nearby but there is a scope of improvements by applying some more fine tuning.

Classifiers	With Fine Tuning						Without Fine Tuning	
	Dataest1			Dataset2			Dataset 1	Dataset 2
Train-Test Ratio	70-30	80-20	90-10	70-30	80-20	90-10	Cross -10	
SVM	93.6	95.7	88	93.9	93.8	94.2	96.7	98.0
NB	95.2	97.9	88	NA	NA	NA	NA	NA
K-NN	88.7	91.5	88	91.6	92.1	92.5	96.3	94.0
D-Tree	90.3	89.4	84	98.8	98.9	98.8	93.5	99.2
NN	94.9	95.3	99.2	93.6	94.1	97.1	NA	NA

Table1: shows the results with different classifies on two different datasets

5. CONCLUSION

From experiments performed in last section we can observed that Performance of support vector machine classifier was highest i.e. 96.7 while we have not performed any fine tuning with dataset 1.Similarly, performance of neural network model was highest i.e. 99.14 while we performed some fine tuning with training testing ratio, goal, learning rate and other parameters with dataset 1. Performance of decision tree classifier was highest i.e. 99.20 while we have not performed any fine tuning with dataset 2. And performance of decision tree classifier was highest i.e. 98.98 while we performed some fine tuning with training testing ratio. From above observations we recommended the use of traditional neural network model for dataset 1 (5 features dataset) and for second dataset we recommend the use of decision tree classifiers. Also the use of neural network model is suggested for some other datasets because there are lots of possibilities of performing fine tuning in neural network in order to increase the overall performance.

REFERENCES

[1] The American Academy of Otolaryngology—Head and Neck Surgery (AAO-HNS). http://www.entnet.org/HealthInformation/Thyroid-Disorders.cfm. Accessed June 2012.

[2]. S.B Patel, P. K Yadav, Dr. D. P.Shukla," Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques", (IOSR-JAVS), e-ISSN: 2319- 2380, p-ISSN: 2319- 2372. Volume 4, Issue 2 (Jul. - Aug. 2013), Pg.no 61-64.

[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 inmedicine. In Proceedings of Journal Medical System 2002.

[5] Polat, K., Sahan, S., Günes, S. A novel hybrid method based on arti?cial immune recognition system (AIRS) with fuzzy weighted pre- processing for thyroid disease diagnosis. Expert Systems with Applications, 32, 1141–1147 (2007).

[6] Jang, J.S.R. ANFIS: Adaptive-Network-Based Fuzzy Inference System. IEEE Transactions on Systems, Man, and Cybernetics.23(3), 665-685 (1993).

[7] Jang, J.S.R., Sun, C.T., Mizutani, E. Neuro-Fuzzy and soft computing. Prentice-Hall: Englewood Cliffs, NJ (1997)

[8] C. Senol , T. Yildirim, "Thyroid and Breast Cancer Disease Diagnosis using Fuzzy-Neural Networks", IEEE, Electric and Electronics Engineering, ELECO 2009, International Conference on 5-8 Nov. 2009, Bursa, pp. II-390 - II-393, (2009).

[9] F. Temurtas, "A Comparative Study On Thyroid Disease Diagnosis Using Neural Networks", Expert Systems With Applications, vol. 36, pp. 944-949, (2009).

[10] F. Saiti, A. Naini, M. Aliyari, M. Teshnehlab, "Thyroid Disease Diagnosis Based on Genetic Algorithms using PNN and SVM", 3rd International Conference on Bioinformatics and Biomedical Engineering (ICBBE 2009), pp. 1-4, (2009).

[11] S. Kamruzzaman, A. Hasan, Ab. Siddiquee and Md. Mazumder, "Medical Diagnosis Using Neural Network", 3rd International Conference on Electrical & Computer Engineering (ICECE 2004), 28-30 Dec. 2004, Dhaka, Bangladesh, (2004).

[12]. Muhammad Anjum Qureshi, Kubilay Eksioglu, "Expert Advice Ensemble for Thyroid Disease Diagnosis", IEEE, 2017.

[13] Jamil Ahmed, M. Abdul Rehman Soomrani," TDTD: Thyroid Disease Type Diagnostics", IEEE, 2016.

[14]. Ahmad Taher Azar , Aboul Ella Hassanien, "Expert System Based On Neural-Fuzzy Rules for Thyroid Diseases Diagnosis", IEEE, 2018.

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

[16] Banu, G. R. (2016). Predicting thyroid disease using linear discriminant analysis (LDA) data mining technique. *Commun. Appl. Electron.(CAE)*, *4*, 4-6.

[17] Ahmad, W., Huang, L., Ahmad, A., Shah, F., & Iqbal, A. (2017). Thyroid diseases forecasting using a hybrid decision support system based on ANFIS, k-NN and information gain method. *J Appl Environ Biol Sci*, *7*, 78-85.

[18] S.Umadevi, Dr.K.S.JeenMarseline "Applying Classification Algorithms to Predict Thyroid Disease" International Journal of Engineering Science and Computing, October 2017 15118 – 15122.

[19] Rastogi, M. V., & LaFranchi, S. H. (2010). Congenital hypothyroidism. Orphanet journal of rare diseases, 5(1), 17.