ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A NEURO COMPUTING FRAME WORK FOR THYROID DISEASE DIAGNOSIS USING MACHINE LEARNING TECHNIQUES

SHAIK RAZIA and M.R.NARASINGARAO
Research Scholar, Department of CSE, K L University, Vaddeswaram, Guntur, Andhra Pradesh, INDIA
Professor, Department of CSE, K L University, Vaddeswaram, Guntur, Andhra Pradesh, INDIA
Email: razia28sk@gmail.com, ramanarasingarao@kluniversity.in

ABSTRACT

Thyroid is a disease, managing of which a difficult proposition in a clinical set up. Many researchers have developed different machine learning techniques to diagnose the disease. Different models had been developed using different algorithms for prediction of the disease. In this proposed research, an integrated model using two neuronal models (SOM and LVQ) have been developed for the purpose. The unlabeled dataset consisting of 215 instances has been gathered from UCI repository. Five inputs have been considered for the model. They are Triiodothyronine, serum Thyroxin, Total Serum Triiodothyronine, TSH, Max TSH and the output is considered as status which has values 1(hyper),2(hypo) and 3(Normal). The integrated model has been developed using competitive learning algorithm along with vector quantization algorithm. The outcome of the integrated model has been compared with decision tree model in predicting the disease and it is observed that the integrated model outperformed the decision tree model with respect to the accuracy of the network. The integrated model has been developed with .net technology and the decision tree model is generated using statistical programming language using R.

Keywords: *Self Organized Maps, Linear Vector Quantization, Triiodothyronine, Serum Thyroxin, Serum Triiodothyronine, Thyroid Stimulating Hormone, Hyper Thyroid, Hypothyroid, Decision Tree.*

1. INTRODUCTION

Among all organs, Thyroid gland is a most significant organ of the endocrine system and it has a weight of 15 -20 gr[1]. It is a butter fly shaped thyroid gland, present below the voice box of the human body. Thyroid gland releases two primary hormones namely Triiodothyronine(T3) and Thyroxine (T4) which controls the total body metabolism [2]. A thyroid disorder is one where a thyroid gland is a defective one or without a thyroid gland. Thyroid disorders can cause the thyroid gland to become overactive (hyperthyroidism) or underactive (hypothyroidism) [3]. The pituitary gland releases a hormone called Thyroid Stimulating Hormone (TSH) located at the base of the brain. It secretes thyroid stimulating hormone (TSH), which controls the thyroid function. To test the functionality of thyroid gland, it is a measure of the Thyroid Simulating Hormone and Triiodothyronine (T3), Thyroxine (T4) levels in the blood sample. Hyper or hypothyroidism is considered to diagnose the thyroid disease [4]. This Paper presents a framework constituting SOM and LVQ and the output of which is compared against Decision tree model using statistical programming language using R.

A self-organizing map (SOM) or self-organizing feature map(SOFM) is a kind of Artificial Neural Network (ANN) that is prepared utilizing unsupervised learning to create a low-dimensional (normally two-dimensional), representation of the input space of the training samples. Self-organizing maps vary from other Artificial Neural Network (ANN) as they apply destructive learning and in the sense that they use a neighborhood function to preserve the topological properties of the input space. Decision Trees are the most famous designs broadly utilized as a part of data mining [5]. At initial, a root node is assigned by utilizing a test. At that point, the estimation of related test characteristic parts the information set and, the procedure is rehashed until the decided ceasing measure is given. Toward the finish of the tree, every node is named as leaf node. Every leaf node means the class. Likewise, every branch demonstrates a way characterized as a choice run the show. The arrangement is taken care of by utilizing every decision manages for another example [6]. As an outline, the decision tree structures comprise of a root node, branches, inward nodes, and leaf-nodes.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2. RELATED WORK

Arpneek Kaur et al. have used Multilayer back propagation network and self organized map for diagnosis of thyroid disease. Weka tool has been used for the purpose. Thyroid data set is used for predicting the disease. The performance of BPN and SOM networks were found by varying the number of neurons present in the hidden layer of the network and also by the percentage of training data [7]. Yilmaz Kaya et.al. have developed a Extreme Learning Model which is a single hidden layer feed forward neural network and is trained with gradient based learning algorithm. This extreme learning model was used in the diagnosis of thyroid disorders by performing classification and the accuracy of the classification was found to be 96.75%. 70% of the samples were used for training and 30% samples were used for testing purpose [8]. Rajkumar Nallamuth et al. have developed various classification models for diagnosis of thyroid disease. The classification models include C4.5, Multilayer Perceptron, and radial basis function networks. It is observed that MLP has performed well compared to other classifiers [9]. Md.Dendi Maysanjaya et al. have used MLP model using back propagation algorithm and the results of which have been compared with WEKA tool and RBF Network. It is found that MLP and RBF have given much accuracy in predicting the thyroid disease [10]. The authors Kenji Hoshi and Junko Kawakami have developed a Bayesian regularized neural network and a self organized map to predict hyperthyroid and hypothyroid using linear discriminate analysis. For this, they have used hormones related thyroid predicting the disease [11]. Similarly Jasdeepsinghbhalla and Anmol agarwal have developed hybrid neural networks for medical diagnosis using scaled conjugate gradient back-propagation and Marquardt back propagation algorithm. The input to the model is a thyroid dataset and basing on this dataset they have compared the performance of the models and found the efficiency of the prediction of ANN's in medical diagnosis [12]. Hasan makas et al. developed seven distinct sorts of Neural Networks keeping in mind the end goal to recognize more strong and dependable systems for diagnosing the thyroid illness. They have utilized swarm optimization and ant bee colony algorithm for training and testing the networks for diagnosis of thyroid disease [13].

3. METHODOLOGY

Supervised learning is an information mining undertaking of inferring a function from named training information. The training information comprise of an arrangement of preparing illustrations. In managed adapting, every case is a couple comprising of an information input object (commonly a vector) and the desired output value (additionally called the supervisory flag). A supervised learning calculation investigates the training information and produces an indirect function, which can be utilized for mapping new illustrations. An ideal improvement will take into account the calculation to effectively decide the class names for unseen cases. This requires the taking in calculation to sum up from the training information to hidden circumstances in a "sensible" manner. In Unsupervised Learning, the preparation of the system is completely information driven and no objective outcomes for the information vectors are given. An ANN of the unsupervised learning sort, for example, the self organizing map can be utilized for bunching the information and discover highlights natural to the issue.

In Information mining or even in information science world, the issue of an unsupervised learning task is attempting to discover hidden structure in unlabeled information. Since the cases given to the model are unlabeled, there is no mistake or reward flag to assess a conceivable arrangement.

In this research, a framework using self organized map (SOM) along with Learning Vector Quantization has been developed. The unlabelled thyroid data of about 215 different patients is obtained from a clinic and is used to train the SOM network using competitive learning algorithm. The competitive learning algorithm divides the dataset into three clusters where each cluster contains the examples of the dataset. The example in each cluster has been labeled as 1, 2 and 3thereby the unlabelled dataset gets converted into a labeled dataset. Each input pattern is a vector of 5 attributes and the output is in the form of a class label '1' ,'2' or '3' which tells whether the patient is diagnosed as having hyper ,hypo and normal. The labeled dataset which is provided by SOM had been given as input to the LVQ network where the network is trained and tested with a supervised algorithm called LVQ algorithm. The dataset consists of 215 patients including the class labels which were given by SOM network had been used in this research. The prediction of LVQ network is compared with the labeled dataset (provided by SOM) in order to ascertain the performance of the LVQ network over SOM although both have been considered together as an integrated model. During testing phase, 50 samples were considered as test dataset for the LVQ network and accuracy of the hybrid neural network is calculated depending on the misclassification results between SOM and LVQ class labels. The performance of the integrated model is compared against the performance of the Decision tree model and the analysis of the decision tree model has been done using a statistical programming language in R [14].

4. DECISION TREE ALGORITHMS

4.1 Self-Organizing Map (SOM) Neural Networks

Self organizing neural networks is an unsupervised neural system comprises of two layers, one layer is the input layer and other layer is an output layer from where the output neuron generate outputs once we give input from the input layer[5]. SOM network uses the competitive learning algorithm to generate the output. The competitive learning algorithm make the network to have a competition among the nodes of the output layer and only one neuron wins the competition. The output of this winning neuron is the output generated by the network.
The steps of the learning algorithm are summarized as follows.

Step1: Initialize all weights to small random values. Set a value for the initial learning rate α and a value for the neighborhood

Step2: Choose an input pattern 'x' from the input data set

Step3: Select the winning unit 'c' (The index of the best matching output unit) such that the performance index I given by the Euclidean distance from x to w_{ij} is minimized.

$I = \| x - W_c \| = \min \| x - W_{ij} \|$

Step4: Update the weights according to the global network updating phase from iteration k to iteration k+1 as : $W_{ij}(k+1) = \{W_{ij}(k)+\alpha(k)[x - W_{ij}(k)]$ if $(i,j) \varepsilon Nc(k)$ $W_{ij}'(k)$ if $(i,j) \notin Nc(k)$
Where α(k) is the adaptive learning rate and Nc(k) is the neighborhood of unit 'c' at iteration k.

Step5: The learning rate and neighborhood are decreased at every iteration according to appropriate scheme. For instance, Kohonen suggested a shrinking function in the form of $\alpha(k)=\alpha(0)[1-k/T]$ with T being the total number of training cycles and α(0) the starting learning rate bounded by one. As for the neighborhood, several researchers suggested an initial region with the size of the half of the output grid which is shrinking according to an exponentially decaying behavior.

Step6: The learning scheme continues until a sufficient number of iterations has been reached or until each output reaches a threshold of sensitivity with respect to a portion of the input space.

4.2 Linear Vector Quantization

Direct Vector Quantization is one case of focused learning. The objective here is to have the system find structure in the information by discovering how data is clustered. The outcomes can be utilized for information encoding and pressure. One such strategy for doing this is called vector quantization. The vector quantization calculation is as per the following. In Linear Vector Quantization, we expect there is a code book, which is described by an arrangement of M model vectors. M is picked by the client and the hidden model vectors are picked subjectively. A data has a place with the cluster i, in the event that i is the list of the closest model (closest in the feeling of the normal Euclidean partition). This is a supervised version of vector quantization. Classes are predefined and we have an arrangement of named information. The objective is to decide an arrangement of models the best represent to each class.

Algorithm:

Step1: Pick the quantity of clusters M

Step2: Initialize the models $w_{*1}, \ldots w_{*m}$ (one basic v strategy for doing this is to randomly pick M directions from the data)

Step3: Reputedly ceasing measure is fulfilled.

Step 4: At any rate pick an information x
Determine the "winning" node k by finding the model vector that performs $|w*_k - x| \leq |w*_i - x| \quad (\forall i)$

Step5: if the models are standardized, this is proportional to boosting w_{*i} x Upgrade just the winning model weights as per w_k (new) $=w_k$ (old) $+\mu(x-w_k$ (old)). This is known as a standard competitive learning principle.

5. RESULTS

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Given below is a table consisting of inputs to identify the thyroid disease. Status 1 and 2 indicate hyperthyroidism and hypothyroidism respectively. The inputs T3,total serum thyroxin, total serum triiodothyronine, TSH and Max TSH have been considered as inputs to the SOM network and each example has been classified into three clusters namely 1 ,2 and 3 respectively thus enabling the unsupervised dataset into a supervised dataset. This clustering has been done using competitive learning algorithm.

In this integrated framework, we have labeled the entire dataset by SOM network using clustering algorithm namely competitive learning algorithm. The labeled dataset has been given as input to the LVQ network using vector quantization algorithm. The table 1 given below is a labeled dataset for 100 samples only, which had been given as input to LVQ network for training and testing using supervised learning.

Given below in table1 consisting of inputs to identify the thyroid disease. Status 1 and 2 indicate hyperthyroidism and hypothyroidism respectively. The inputs T3,total serum thyroxin, total serum triiodothyronine, TSH and Max TSH have been considered as inputs to the SOM network and each example has been classified into three clusters namely 1 ,2 and 3 respectively thus enabling the unsupervised dataset into a supervised dataset. This clustering has been done using competitive learning algorithm.

In this integrated framework, we have labeled the entire dataset by SOM network using clustering algorithm namely competitive learning algorithm. The labeled dataset has been given as input to the LVQ network using vector quantization algorithm. The table 1 given below is a labeled dataset for 100 samples only, which had been given as input to LVQ network for training and testing using supervised learning.

Table 1: The clustering results given by the SOM network for 100 samples

T3	Total Serum Thyroxin	Total Serum Triiodothyronine	TSH	MAX TSH	SOM Status
107	10.1	2.2	0.9	2.7	1
116	10	1.7	1.5	4.3	1
117	9.2	1.9	1.5	6.8	1
106	6.7	1.5	1.2	3.9	2
92	11.1	2	0.7	-0.2	1
110	10.4	1.8	1	2.3	2
120	8.4	1.1	1.4	1.4	1
116	11.1	2	1.2	2.3	1
110	7.8	1.9	2.1	6.4	2
113	9.9	3.1	2	5.9	1
134	1.9	0.6	18.4	8.2	1
108	7.1	1.3	1.6	2.2	2
105	5.7	1	0.9	0.9	1
139	16.4	3.8	1.1	-0.2	2
109	8.4	2.1	1.1	3.6	1
111	8.4	1.5	0.8	1.2	1
123	1.9	0.3	22.8	22.2	1
127	7.7	1.8	1.9	6.4	2
108	6.5	1	0.9	1.5	1
112	2.6	0.7	41	19	2
101	6.7	1.3	1	5.7	3
111	16	2.1	0.9	-0.1	1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

106	9.4	1.5	0.8	0.5	3
130	9.5	1.7	0.4	3.2	1
100	10.5	2.4	0.9	1.9	3
126	0.5	0.2	12.2	8.8	1
115	10.6	0.8	2.1	4.6	3
130	10	1.6	0.9	4.6	1
105	12	3.3	1.1	0	3
127	12.9	2.4	1.4	0.6	1
122	9.7	1.6	0.9	2.2	1
125	2.3	0.9	16.5	9.5	1
106	8.9	0.7	1	1.5	3
102	8.4	1.5	0.8	1.4	1
93	8.9	1.5	0.8	2.7	1
120	10.4	2.1	1.1	1.8	3
106	11.3	1.8	0.9	1	1
120	6.8	2.1	10.4	38.6	2
123	8.1	2.3	1	2.3	3
107	8.4	1.8	1.5	0.8	1
109	10	1.3	1.8	4.3	1
144	22.3	3.3	1.3	0.6	3
116	11.9	1.8	1.9	1.5	1
113	11.5	1.5	1.9	2.9	1
76	25.3	4.5	1.2	-0.1	1
110	8.7	1.9	1.6	4.4	2
103	8.1	1.4	0.5	3.8	1
101	7.1	2.2	0.8	2.2	1
115	10.4	1.8	1.6	2	2
110	15.2	19	0.7	-0.2	1
115	6.3	1.2	4.7	14.4	2
101	6.3	1.5	0.9	2.9	1
106	4.2	1.2	1.6	1.4	1
110	11.3	2.3	0.9	3.3	1
113	17.2	1.8	1	0	1
114	6.7	1.5	1	3.5	1
119	10.6	2.1	1.3	1.1	1
115	7.1	1.3	1.3	2	1
108	3.5	0.6	1.7	1.4	1
111	9.1	1.7	1.2	4.1	1
120	3	2.5	1.2	4.5	1
117	7.8	2	1	3.9	1
97	4.7	1.1	2.1	12.6	2
90	8.1	1.6	1.4	1.1	1

117	12.2	1.9	1.2	3.9	2
117	11	1.4	1.5	2.1	1
113	9	2	1.8	1.6	1
102	5.3	1.4	1.3	6.7	1
102	6.6	1.2	1.4	1.3	1
88	16.5	4.9	0.8	0.1	1
103	5.1	1.4	1.2	5	2
111	11.9	2.3	0.9	3.8	1
104	6.1	1.8	0.5	0.8	2
118	6.5	1.3	1.7	11.5	1
126	10.4	1.7	1.2	3.5	1
114	7.5	1.1	1.6	4.4	2
122	11.8	2.7	1.7	2.3	1
105	8.1	2	1.9	-0.5	2
109	7.6	1.3	2.2	1.9	1
105	9.5	1.8	1.6	3.6	2
139	4.2	0.7	4.3	6.3	3
79	19	5.5	0.9	0.3	2
118	12.2	1.5	1	2.3	3
94	7.5	1.2	1.3	4.4	2
98	5.7	0.4	1.3	2.8	1
112	6.5	1.2	1.2	2	3
97	15.1	1.8	1.2	-0.2	2
106	13.4	3	1.1	0	1
121	10.1	2.4	0.8	3	3
110	9.2	1.6	1.5	6.4	1
129	11.9	2.7	1.2	3	3
121	13.5	1.5	1.6	0.1	1
119	3.8	1.1	23	5.7	1
106	9.4	2.2	1.5	0	2
107	13	1.1	0.9	3.1	1
99	17.5	1.9	1.4	0.3	1
109	5.3	1.6	1.4	1.5	2
94	20.5	1.8	1.4	-0.5	1
98	9.1	1.4	1.9	-0.3	2
111	7.8	2	1.8	4.1	1

LVQ TRAINING

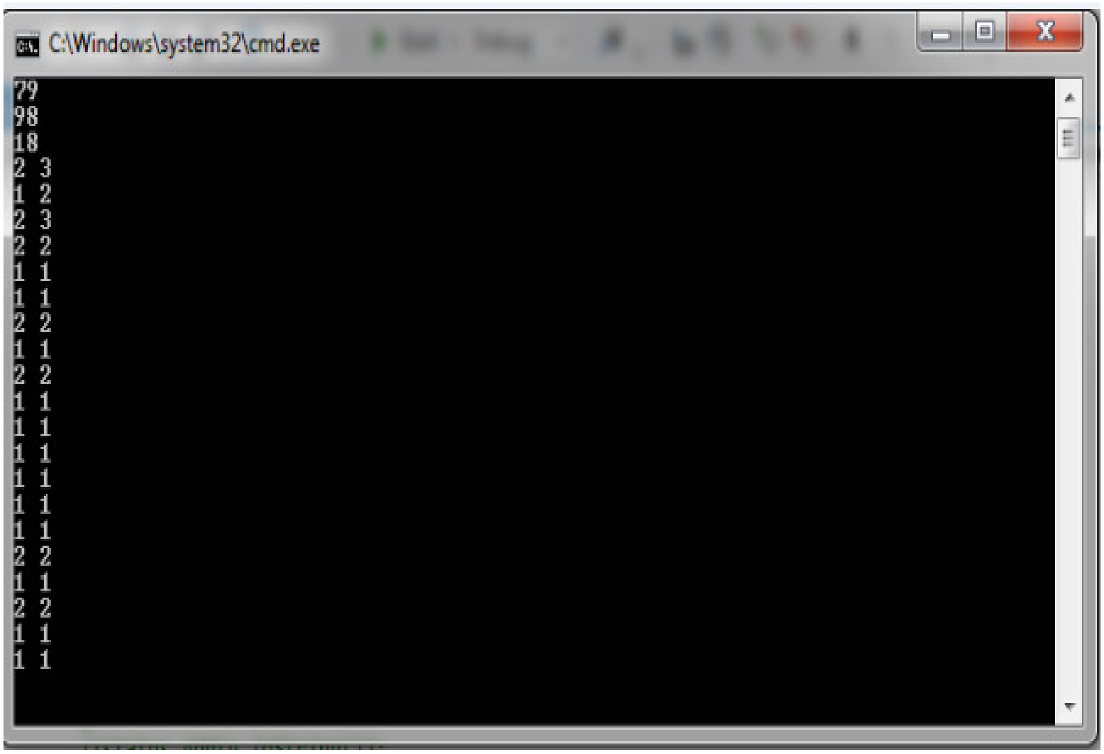
In table no: 2 the comparison of the LVQ network with SOM network. It is observed that 85% accuracy has been achieved by LVQ network in-predicting the class labels which were previously labeled by SOM network. Similarly, the integrated model has outperformed the decision tree model (75%) which is very much less compared with integrated model.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

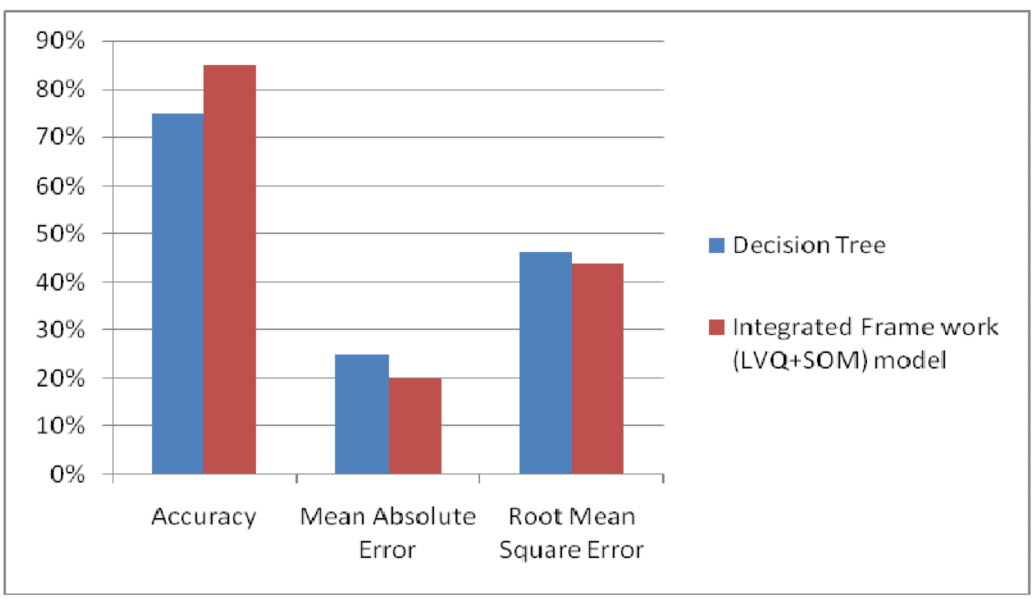
Table 2: The Testing results shown by the LVQ network

T3	Total Serum Thyroxin	Total Serum Triiodothyronine	TSH	MAX TSH	SOM Status	LVQ Status	Decision Tree Status
98	8.6	1.6	1.6	6	3	2	2
109	12.4	2.3	1.7	0.8	2	1	2
67	23.3	7.4	7.8	-0.6	3	2	2
95	11.1	2.7	1.6	-0.3	2	2	1
117	6.7	2.2	1.8	6.7	1	1	1
115	15.3	2.3	2	2	1	1	1
91	8	1.7	2.1	4.6	2	2	2
103	8.5	1.8	1.9	1.1	1	1	1
89	23.8	5.4	0.5	0.1	2	2	1
126	9.4	2.3	1	4	1	1	1
113	8.5	1.8	0.8	0.5	1	1	1
109	9.7	1.4	1.1	2.1	1	1	1
119	12.9	1.5	1.3	3.6	1	1	1
101	7.1	1.6	1.5	1.6	1	1	1
108	10.4	2.1	1.3	2.4	1	1	1
89	14.3	4.1	0.5	0.2	2	2	1
118	10.5	2.1	0.7	3.5	1	1	1
97	7.8	1.3	1.2	0.9	2	2	2
113	11.1	1.7	0.8	2.3	1	1	2
112	9.5	2	1.2	0.7	1	1	1

Thyroid results using decision trees and integrated model



The Performance Of The Decision Tree Model With The Integrated Model Is Given Below



The Graph Given Above Gives Information About The Accuracy Of The Integrated Frame Work Model Over Decision Tree Model.

Table 3: Performance Accuracy Of The Models Decision Tree And Integrated Framework Model.

Classifier	Algorithm	Accuracy	Mean Absolute Error	Root Mean Square Error
Decision Tree	Decision Stump	75%	0.25	0.46
Integrated Frame work (LVQ+SOM) model	LVQ and SOM Algorithm	85%	0.20	0.44

6. DISCUSSION

Neural Networks have been successfully applied in different fields of the medical domain especially in the diagnosis of disease. In this proposed research, an integrated model using SOM and LVQ has been developed and the objective of which is to classify the disease as Hyperthyroid, Hypothyroid and Normal and has performed well with other models like decision trees. The integrated model has been used for clustering by SOM network and has been given as input to the LVQ network within the integrated model. Although neural network have been used extensively in medical diagnosis, but to our knowledge, this is the first time this kind of integrated model had been used for the diagnosis of thyroid disease [15]. In fact, the development of our integrated model is a practical tool to predict thyroid disease based on the limited dataset that is available with us. The model can further be enhanced to any desired level by increasing the number of inputs and outputs and dynamic data can be generated as more data can be fed to it. In a nutshell, not only we have developed a prototype integrated frame work to diagnose the thyroid disease but also acts as a decision maker for diagnosing the thyroid disease.

7. CONCLUSION

In summary, we developed an integrated model using the available information and the results of which have been consistent with the domain experts.

REFERENCES

[1] http://cdn.intechopen.com/pdfs-wm/31300.pdf
[2] http://www.rightdiagnosis.com/t/thyroid/intro.htm
[3] Nadeau, Andre. "Subclinical Thyroid Disease: A laboratory Finding of Clinical relevance." Canadian Journal of CME 13.7 (2001): 89-98.
[4] Shroff, Shanu, et al. "Thyroid disease diagnosis: A survey." Intelligent Systems and Control (ISCO), 2015 IEEE 9th International Conference on. IEEE, 2015.
[5] Turanoglu-Bekar, Ebru, Gozde Ulutagay, and Suzan Kantarcı-Savas. "Classification of Thyroid Disease by Using Data Mining Models: A Comparison of Decision Tree Algorithms." (2016).
[6] J. J. Oliver, D. Hand, Averaging Over Decision Stumps, in Machine Learning, In Proceedings of European Conference on Machine Learning (ECML), Catania, Italy, (1994). http://dx.doi.org/10.1023/10.1007/3-540-57868-4_61[13]
[7] Arpneek Kaur, Narpat Singh, and Abhishek Bahrdwaj, 'A Comparison of Supervised Multilayer Back Propagation and Unsupervised Self Organizing Maps for the Diagnosis of Thyroid Disease', International Journal of Computer Applications, 82 (2013)
[8] KAYA YILMAZ. "A Fast Intelligent Diagnosis System for Thyroid Diseases based on extreme learning machine." ANADOLU University Journal of Science and Technology. A Applied Sciences and Engineering 15.1 (2014): 41-49.
[9] Rajkumar, Nallamuth, and Jacanathan Palanichamy. "Optimized construction of various classification models for the diagnosis of thyroid problems in human beings." Kuwait Journal of Science 42.2 (2015).
[10] Maysanjaya, I. Md Dendi, Hanung Adi Nugroho, and Noor Akhmad Setiawan. "A comparison of classification methods on diagnosis of thyroid diseases."Intelligent Technology and Its Applications (ISITIA), 2015 International Seminar on. IEEE, 2015.
[11] Hoshi, Kenji, et al. "An analysis of thyroid function diagnosis using Bayesian-type and SOM-type neural networks." Chemical and pharmaceutical bulletin 53.12 (2005): 1570-1574.
[12] Bhalla, Jasdeep Singh, and Anmol Aggarwal. "A novel method for medical disease diagnosis using artificial neural networks based on

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

backpropagation algorithm." Confluence 2013: The Next Generation Information Technology Summit (4th International Conference). IET, 2013.

[13] Makas, Hasan, and Nejat Yumusak. "A comprehensive study on thyroid diagnosis by neural networks and swarm intelligence." Electronics, Computer and Computation (ICECCO), 2013 International Conference on. IEEE, 2013.

[14] Ihaka, Ross, and Robert Gentleman. "R: a language for data analysis and graphics." Journal of computational and graphical statistics 5.3 (1996): 299-314.

[15] Shaik Razia, M R Narasingarao, " Machine Learning Techniques for thyroid disease diagnosis-A Review", Indian Journal of Science and Technology, Vol:9, Issue:28, Pg.No:1-9, July,2016.