Diagnosis of Thyroid Disease Using Some Machine Learning Technique & Comparatively Analysis

[1]Ankita Soni, [2]Ram Nivas Giri

[1] M. Tech Scholar, Department of Computer Science & Engineering, Raipur Institute of Technology, Raipur (C.G), India
[2]Asst. Professor, Department of Computer Science & Engineering, Raipur Institute of Technology, Raipur (C.G), India

Abstract: The thyroid gland is one of the most important organ in our body. It secretes thyroid hormones, which are responsible for controlling metabolism. The less secretion hormone causes hypothyroidism and much secretion of thyroid causes hyperthyroidism. Therefore, in medical disease classification, choice of classifier plays a vital role. This paper describes the diagnosis of thyroid disorders using some machine learning techniques like Decision Tree, Support Vector Machine and Neural Networks. The performance evaluation of this system is estimated by using classification accuracy and k-fold cross-validation. Finally, results are compared based on confusion matrix, which shows the average accuracy of 97%. The thyroid data set is obtained from UCI machine learning repository. For simulation, MATLAB has been used.

***Keywords* – Thyroid disease, KNN, Decision Tree, Quadratic Discriminant and Support Vector Machine.**

I. Introduction

In the Medical Science field, the most monotonous and challenging task is to provide disease diagnosis at early stage with higher accuracy. Thyroid disorder is a major public health problem. Millions of people in the world have thyroid disorders especially women. Most of them have undiagnosed thyroid diseases, which becomes a major concern. An abnormality or imbalance in production of thyroid hormones can cause a host of problems, from fatigue and depression to weight gain leading to thyroid problems. The thyroid is a butterfly-shaped gland located in the front of the neck just below the Adams apple formed by two wings (lobes) represented by the left and right thyroid lobes, it's a biggest gland in the neck which is placed in the anterior neck. This gland produces thyroid hormones. These hormones help to regulate the body's metabolism and effects processes, such as growth and other important functions of the body. The two most important hormones are thyroxine (T4) and triiodothyronine (T3). The hormone with the most biological power is actually T3. Once released from the gland into the blood, a large amount of T4 is converted to T3 - the active hormone that affects the metabolism of cells throughout the body. Thyroid disorders can cause the thyroid gland to become overactive (hyperthyroidism) or underactive (hypothyroidism). There are two most common problems of thyroid disorder or thyroid disease. They are Hyperthyroidism – releases too much thyroid hormone into the blood due to over active of thyroid and Hypothyroidism - when the thyroid is not active and releases too low thyroid hormone into the blood [1]

Thyroid diseases are broadly divided into three types (i) Hyperthyroid: Increase in the hormone production can cause hyperthyroidism. Hyperthyroidism occurs when the gland produces excess hormones. The symptoms that indicate the presence of hyperthyroidism includes loss of weight, high blood pressure, nervousness, increase in heart rate, an increased sweating, swelling in your neck, frequent bowel movements, shorter menstrual periods and trembling hands. (ii) Hypothyroid: Decrease in the hormone production can cause hypothyroidism. In medical field, Hypothyroidism is a condition that the thyroid gland does not produce enough hormones. Inflammation and damage to the gland causes hypothyroidism. (iii) Thyroiditis: Thyroiditis is a thyroid disease in which the patient has an inflammation of her thyroid gland.

The symptoms of Thyroiditis includes fatigue, depression, cold intolerance, weight gain, dry skin and hair, muscles, constipation, decreased concentration and sleepiness, leg swelling, puggy eyes, severe symptoms include a slow heart rate, low body temperature, heart failure and coma.

Some most commonly used classification techniques for thyroid detection from various available datasets are KNN (k Nearest Neighbor), Decision Tree, Quadratic Discriminate and Support Vector Machine.

II. RELATED WORK

Nikita Singh and Alka Jindal, have concluded that SVM is far much better classifier as compared to KNN and Bayesian. Accuracy of SVM is about 84.62%. KNN found the nearest neighborhood automatically [2].

Mary C. Frates et al. have provided in her paper US features associated with thyroid cancer. They have also suggested that which nodules should be subjected to US-guided fine needle aspiration and which thyroid nodules need not be subjected to fine-needle aspiration. SVM is the best available machine learning algorithms in classifying high-dimensional data sets [3].

Farhad Soleimanian Gharehchopogh et al**.** considered a Multi-layer Perceptron (MLP) ANN using back propagation learning algorithm to classify Thyroid disease. The accuracy level for thyroid disease is reached to 98.6% and also increase the performance of ANN [25].

Feyzullah Temurtas have used Various data mining techniques like Bayes net, MLP, RBF Network, C4.5, CART, REP tree and decision stump to develop classifier for diagnosis of hypothyroid disease and yielded 99.60%accuracyThe entire work is simulated in WEKA tool [4].

Anupam Skukla et al. proposed the diagnosis of thyroid disorders using Artificial Neural Networks (ANNs). Three ANN algorithms has been used for the diagnosis, they are, the Back propagation algorithm (BPA), the Radial Basis Function (RBF), and the Learning Vector Quantization (LVQ) Networks. The accuracy was obtained for BPA is 92%, for RBF is 80% and for LVQ is 98% [5]. The dataset of thyroid was taken from UCI repository of machine databases [13]

F.Temurtas, realized the diagnosis by multilayer, probabilistic, and learning vector quantization neural networks were implemented on thyroid disease. and the achieved accuracy 92.96 %, 94.43% and 89.79% respectively [6].

Li-Na Li, Ji-Hong Ouyang, Hui-Ling Chen & Da-You Liu developed A CAD system PCA-ELM for assisting the diagnosis of thyroid disease. It was observed that PCA-ELM achieved the highest classification accuracy of 98.1% and mean classification accuracy of 97.73% using 10-fold cross-validation. The experimental outcome showed that PCA-ELM performed bether than PCA-SVM in terms of classification accuracy with shorter run time [7].

G. Rasitha Banu; In this work, Two data mining techniques such as J48 and Decision stump tree are used to classify hypothyroid disease and achieved high accuracy of 99.57% and 95.38% respectively [8].

Ahmad Taher Azar and Aboul Ella Hassanien**”** proposed a method for thyroid disease diagnosis using ANFIS**.** Adaptive Neuro-Fuzzy Classifier with Linguistic Hedges has been used in this paper. The results indicated that the classification accuracy without feature selection was 98.6047% and 97.6744% during training and testing phases, respectively with RMSE of 0.02335. After applying feature selection algorithm, LHNFCSF achieved 100% for all cluster sizes during training phase. However, in the testing phase LHNFCSF achieved 88.3721% using one cluster for each class, 90.6977% using two clusters, 91.8605% using three clusters and 97.6744% using four clusters for each class and 12 fuzzy rules [9].

J. Jacqulin Margret, B. Lakshmipathi and S. Aswani Kumar**,** Various decision tree splitting rules are used and they are Information gain, Gini Index, Likelihood Ration Chi-Squared Statistics, Gain Ratio and Distance measure. From this work, it is clear, that normalized based splitting rules have high accuracy and sensitivity or true positive rate [10].

Zhang GP, Berardi proposed a model to diagnose the thyroid dysfunction using Artificial Neural Network (techniques) models like Te cross validation method, Variable Selection method and Te Regression based method. The result found is the neural network have shown best accuracy result in diagnosing the thyroid dysfunction [11].

Hoshi K, Kawakami J, Kumagai M, Kasahara S, Nishimura N, Nakamura H, Sato K proposed an important classification problem for thyroid function diagnosis by using multivariate analysis and two notable approaches the Bayesian regularized networks and also the Self Organising Map (SOM) [12]

Keles A. proposed a method with expert system for thyroid disease diagnosis. The method used for the thyroid disease called as ESTDD (Expert System Thyroid Disease Diagnosis). ESTDD diagnose with accuracy 95.33% for thyroid disease [14].

Temurtas F. compared the thyroid ailment analysis by using multilayer (NN), Probabilistic (NN) and learning vector quantization (LVQ-NN) neural systems. Result demonstrates that Probabilistic neural system gives the best grouping exactness for thyroid infection finding [15].

Sharpe PK, Solberg HE, Rootwelt K, Yearworth M. examined two sorts of ANN design and evaluated their power in finding of thyroid capacity. The multilayer perceptron which is developed by back propagation algorithm and learning vector quantization have been used for designing and testing procedures [16].

Kousarrizi, Nazari MR, Seiti F, Teshnehlab M. examined a few techniques for highlighting to choice and order for thyroid sickness analysis. They proposed an essential component called choice technique, hereditary calculation for nonlinear streamlining issues. Bolsrer vector machine is utilized as classifier to analyze the thyroid sickness and it is observed that Support Vector outperformed Bolster vector machine. Result found was classification rate is observed to be 98.62% for SVN and which has given precision of 99.6% for preparing and testing [17].

Rouhani M, Mansouri K. compared and studied a few ANN structure on the thyroid ailment finding. To analyze the sickness, RBF, LVQ and SVM system models are utilized for the purpose. It is observed that, RBF Networks gave the best result for determination of thyroid organ capacity. RBF Network outperformed others systems in analyzing the thyroid disease [18].

Shariati, Hanhighi MM. Examined the fuzzy neural network and compared the performance of fuzzy NN and Support Vector Machine for identification of thyroid disease. The author have determined that the fuzzy NN has outperformed the previous classification by taking into consideration of specific parameters like Hyper, Hypo, Sub-clinical hyperthyroid, Sub-clinical hypothyroid and no thyroid prediction of thyroid disease is found to be 95.4% to 99.5% [19]

In [34] Proposed a classification method over support vector machines and probabilistic neural systems for identifying the disease diagnosis. It is observed that SVM has performed superior to other networks.

Dogantekin E, Dogantekin A, Avei D. proposed a technique to diagnose the thyroid disease, using generalized discriminant analysis and wavelet support vector machine system [20].

Aziz SB. developed an NN model for diagnosis of thyroid disease using GA. MATLAB is used for simulation for NN and training and testing was found to be between 96% and 98% [21]

Prerana PS. presented three methods of neural systems for prediction of thyroid sickness. Among these, the authors have proposed the best techniques for determination of thyroid disease and reduce the diagnosis time and increased the accuracy [22].

Banu GR Predicted hypothyroid disease using data mining algorithm called Linear Discriminant Analysis (LDA) to enhance the accuracy. The LDA algorithm gives accuracy of 99.62% with cross validation K=6 [23]

III. METHODOLOGY

This section explains about the step by step procedure of algorithm and various techniques used for this work. The block diagram for this work is shown below in figure 3.1

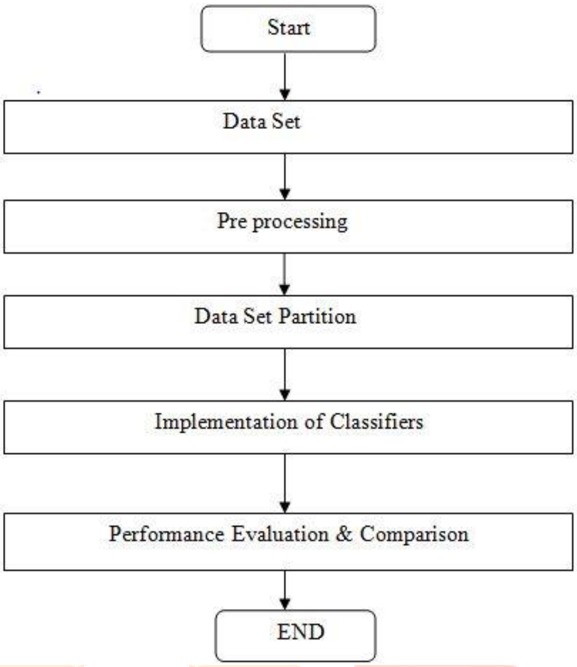

Start
Data Set
Pre processing
Data Set Partition
Implementation of Classifiers
Performance Evaluation & Comparison
END

Figure 3.1: Flowchart of Methodology

3.1 Data Collection

Thyroid data set is collected from UCI repository. The 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. The data set content 215 samples and each sample has 5 features.

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 radioimmuno assay.
4. Basal thyroid-stimulating hormone (TSH) as measured by radioimmuno assay.
5. Maximal absolute difference of TSH value after injection of 200 µg of thyrotropin-releasing hormone as compared to the basal value.

The 150 samples of 215 belong to hyper-function class namely class-1. The 35 samples of 215 belong to hypo function class namely class-2. The 30 samples of 215 belong to normal-function class namely class-3 [4].

3.2 Pre processing

All the null and duplicate values has been removed from dataset.

3.3 Dataset Partition

In data partition stage input data set has to be divided into two sets i.e. training set and testing set, Data partition stage generate two mutually exclusive data set shares no data among each other and both the set having unique content of data set.

3.4 Classifiers

In data partition stage input data set has to be divided into two sets i.e. training set and testing set, Data partition stage generate two mutually exclusive data set shares no data among each other and both the set having unique content of data set.

3.1.1 K-Nearest Neighbor Method: KNN Classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it. Each tuple represents a point in an n-dimensional space. When given an unknown tuple, a knearest neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. Based on these k training tuples are the k "nearest neighbors" of the unknown tuple. The unknown tuple is classified by a majority vote of its neighbors, and gets assigned to the class most common amongst its k-nearest neighbors. When given a training tuple k-Nearest Neighbor simply stores it and waits until it is given a test tuple. Hence it is a "lazy learner" as it stores the training tuples or the "instances", they are also known as "Instance- Based Learners".

3.1.2 Decision Tree: A decision tree is a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test and each leaf node holds a class label. The top most node in a tree is the root node. The construction of decision tree classifiers does not require any domain knowledge and appropriate for exploratory knowledge discovery. Decision tree induction can be used for simple and fast classification. Decision tree algorithms can be used for classification in many application areas such as medicine, astronomy, financial analysis, molecular biology.

3.1.3 QDA (Quadratic Discriminant Analyzer): A quadratic classifier is used in machine learning and statistical classification to separate measurements of two or more classes of objects or events by a quadric surface. It is a more general version of the linear classifier. QDA is closely related to linear discriminant analysis (LDA), where it is assumed that the measurements are normally distributed. Quadratic Discrimination is the general form of Bayesian discrimination.

3.1.4 Quadratic Support Vector Machine: It is a supervised machine learning algorithm that can be employed for both classification and regression purposes. SVMs are more commonly used in classification problems. SVM is a binary classifier that makes its decision by constructing a linear boundary or hyperplane that separates data points of t.he two classes optimally, in Feature Hyperspace.

3.5 Performance Evaluation and comparatively analysis

In this stage to perform analysis of various classifiers and evaluation of their performance, the efficiency of any classifier is moving around how accurately any classifier perform classification. To measure of any classifier some factor like, the specificity, sensitivity, positive and negative predictive and classification accuracy need to be calculated and analyzed. These factors has been calculated with the help of confusion matrix.

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. Following table 3 (real v/s predicted) shows the confusion matrix.

.

Table 3: Definition of Confusion Matrix

Predicted	**Real**	
	Positive	**Negative**
Positive	True positive (TP)	False positive (FP)
Negative	False negative (FN)	True negative (TN)

The entries in the confusion matrix have the following meaning in the context of our study:

- TN is the number of correct prediction that an instance is negative.
- FN is the number of incorrect prediction that an instance is positive.
- FP is the number of incorrect prediction that an instance is negative.
- TP is the number of correct prediction that an instance is positive.

Following factors can be calculate with the help of confusion matrix

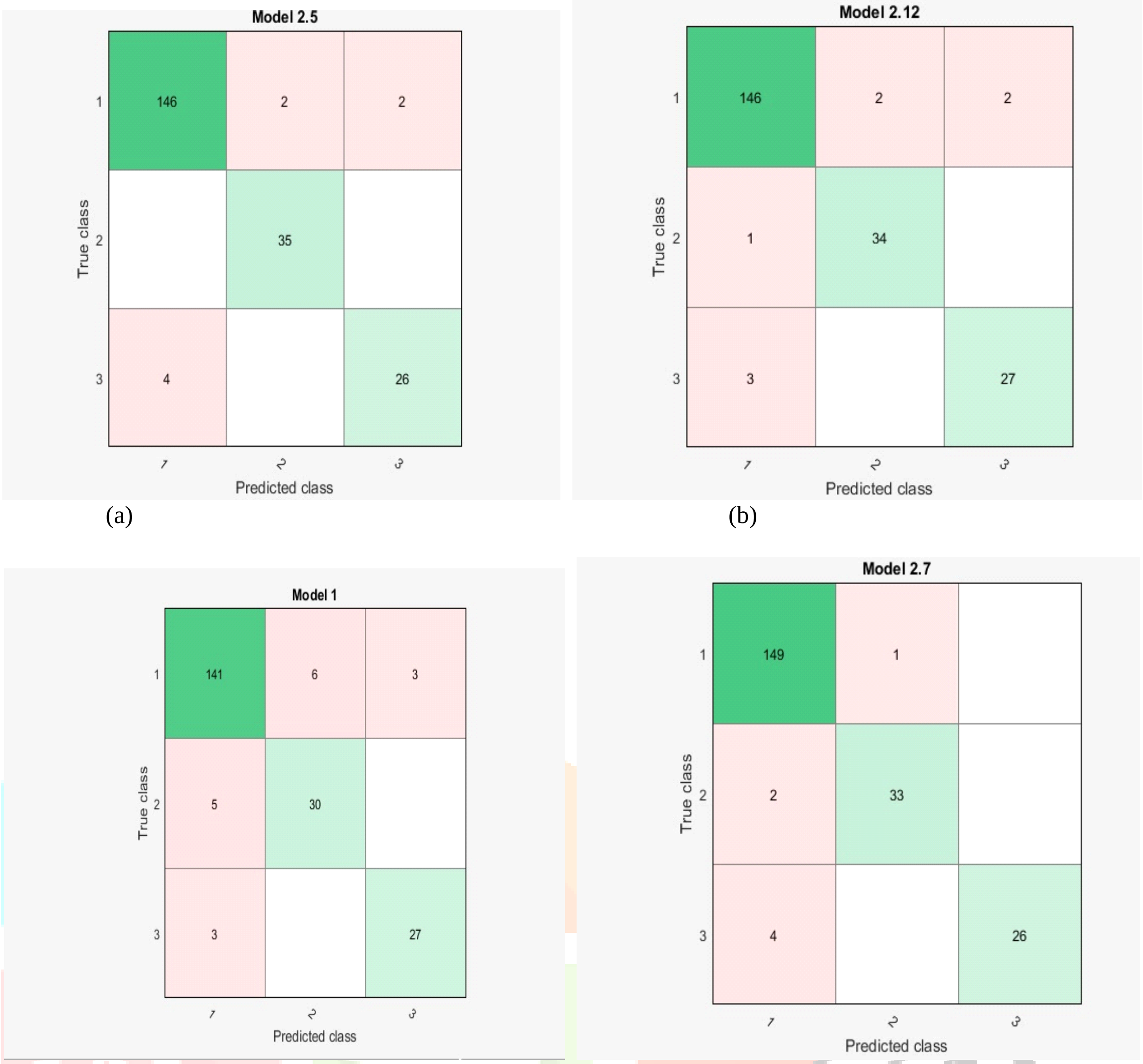
Specificity: It is the capability of the classifier to classify accuretly negative instances (TN) from all the negative instances classified.

$$\text{Specificity (\%)} = [TP / (FP+TN)]*100$$

Accuracy: It is defined as the Sum of All Diagonal Elements in Confusion Matrix/All Matrix Element

IV. RESULT

This section shows the result and graph obtained from this work. The thyroid data set is obtained from UCI machine learning repository has been used for this experiment. It was trained by developing code in MATLAB. To avoid over fitting problems during modeling process, k-fold cross-validation was used for better reliability of test results [25]. In k-fold cross-validation, the original sample is randomly partitioned into k subsamples. A single subsample is retained as the validation data for testing the model, and the remaining k - 1 sub samples are used as training data. The cross-validation process is then repeated k times (the 'folds'), with each of the k subsamples used exactly once as the validation data. The average of the k results gives the validation accuracy of the algorithm [26]. The advantages of k-fold cross validation are that the impact of data dependency is minimized and the reliability of the results can be improved [27].Below is the figure of confusion matrix for each classifiers



(a) (b)

(c) (d)

Figure 4.1: Confusion matrix of (a) Quadratic Discreminant Algorithm (b) K- nearest neighbor Algorithm (c) Decision Tree Algorithm (d) Quadratic Support Vector Machine Algorithm

Table 4.1: Showing Specificity of Different Algorithms

Class/ Algorithm	KNN	Decision Tree	QDA	SVM
Class 1	0.938	0.877	0.938	0.908
Class 2	0.988	0.967	0.988	0.995
Class 3	0.989	0.984	0.989	1.000

The above table is showing the specifity of KNN, Decision tree, Quadratic Discriminant and Support Vector Machine for all three different classes.

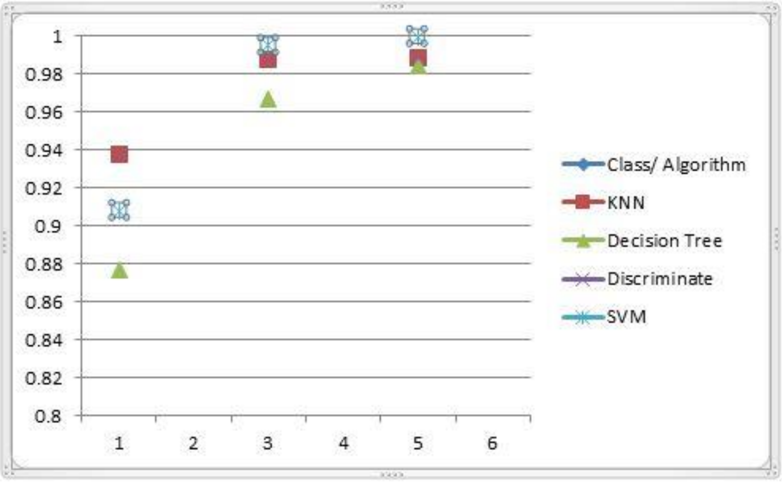


Figure 4.2: Specificity of Different Algorithms

Table 4.2: Showing Accuracy of Different Algorithms

S. No.	Algorithm	Accuracy
1	KNN (k Nearest Neighbor)	96.3
2	Decision Tree	92.1
3	Quadratic Discriminate	96.3
4	Quadratic SVM	96.7

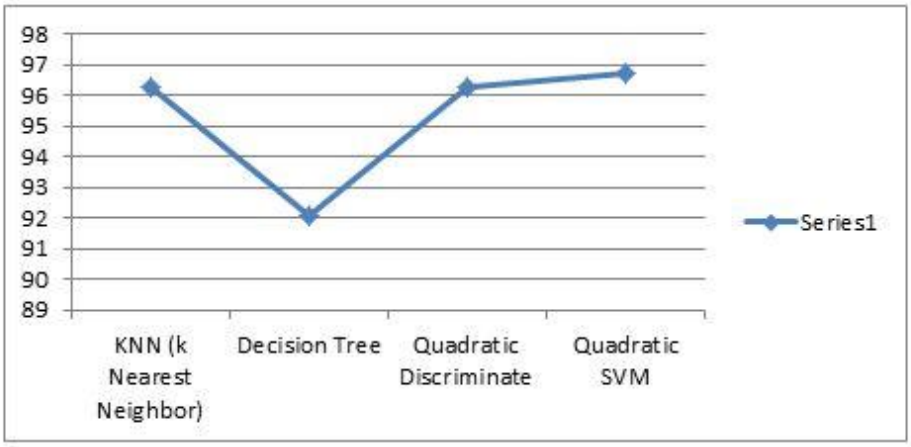


Figure 4.3: Accuracy of Different Algorithms

V. CONCLUSION

Different Researchers have proposed different techniques to predict the thyroid disorder and different kinds of accuracy level as per used techniques but on comparing this technique with all previous methods we can observe that accuracy of quadratic support vector machine is 96.7% which is higher than all other classifiers shown in table 4.1 and figure 4.3.

REFERENCES

[1] Jiawen Han, Micheline Kamber, “DataMining Concepts and Techniques”.

[2] Nikita Singh, Alka Jindal, “A Segmentation Method and Comparison of Classification Methods for Thyroid Ultrasound Images”, International Journal of Computer Applications (0975 – 8887) Volume 50 – No.11, July 2012

[3] Mary C. Frates, Carol B. Benson, J.William Charboneau and Edmund S. “Management of Thyroid Nodules Detected at US: Society of Radiologists in US consensus”, conference statement management of thyroid nodules detected at US Volume 237, Number3Isa IS, Saad Z, Omar S, Osman MK, Ahmad KA, Sakim H. Suitable MLP Network activation functions for breast cancer and thyroid disease diagnosis. 2010 Second International Conference on Computational Intelligence, Modelling and Simulation (CIMSIM); 2010

[4] Shivanee Pandey, Rohit Miri, S. R. Tandan, "Diagnosis and Classification of Hypothyroid Disease Using Data Mining Technique", TJERT, June 2013.Keles, A Expert system for thyroid disease diagnosis. Expert Systems with Applications; 2008

[5] Anupam Shukla, Prabhdeep Kaur, Ritu Tiwari and R.R. Janghel, “Diagnosis of Thyroid disease using Artificial Neural Network”. In Proceedings of IEEE IACC 2009, pages 1016-1020.

[6] Feyzullah Temurtas. “A comparative study on thyroid disease diagnosis using neural networks” Expert Systems with Applications 36 (2009) 944–949.

[7] Li-Na Li,Ji-Hong Ouyang ,Hui-Ling Chen &Da-You Liu. “A Computer Aided Diagnosis System for Thyroid Disease Using Extreme Learning Machine”J Med Syst (2012) 3327–3337.

[8] G. Rasitha Banu “ A Role of decision Tree classification data Mining Technique in Diagnosing Thyroid disease” International Journal of Computer Sciences and EngineeringVolume-4, Issue-11 2016.

[9] Ahmad Taher Azar and Aboul Ella Hassanien “Expert System Based on Neural fuzzy rules for thyroid disease diagnosis”, IEEE

[10] J. Jacqulin Margret, “Diagnosis of Thyroid Disorders using Decision Tree Splitting Rules” International Journal of Computer Applications (0975 – 8887) Volume 44– No.8, April 2012

[11] Zhang GP, Berardi. “An investigation of neural network in thyroid function diagnosis”. Health Care Management Science, 1998;1:29-37

[12] Hoshi K, Kawakami J, Kumagai M, Kasahara S, Nishimura N, Nakamura H, Sato K, “An Analysis of thyroid function diagnosis using Bayesian-type and SOM-type neural networks”. Chemical and Pharmaceutical Bulletin; 2005

[13] UCI repository of machine learning databases,http.//archive. ics. uci. edu/ml/datasets.

[14] Keles A, “Expert System for thyroid disease diagnosis, Expert systems with Applications”; 2008

[15] Temurtas F. “A comparative study on thyroid disease diagnosis using neural networks. Expert system with Applications”; 2009

[16] Sharpe PK, Solberg HE, Rootwelt K, Yearworth M. “Artificial neural network in diagnosis of thyroid function in vitro laboratory test”. Clinical Chemistry; 2009

[17] Kousarrizi, Nazari MR, Seiti F, Teshnehlab M. “An experimental comparative study on thyroid disease diagnosis based on feature subset selection nd classification”. International Journal of Electrical & Computer Science; 2012

[18] Rouhani M, Mansouri K. “Comparison of several ANN architecture on thyroid disease grades diagnosis”. Computer Science and Information Technology – Spring Conference; 2009

[19] Shariati, Hanhighi MM. “Comaparison of ANFIS neural network with several other ANN’s and Support Vector Machine for diagnosis hepatitis and thyroid disease”; 2010 . International Conference on Computer Science systems and Industrial Management Applications (CISM); 2010

[20] Dogantekin E, Dogantekin A, Avei D. “An Expert system based on generalized discriminant analysis and wavelet support vector machine for diagnosis of thyroid disease”. Expert System with Applications; 2011

[21] Aziz SB. “Thyroid disease diagnosis using Generic Algorithm and Neural Network”; 2009

[22] Prerana PS. “Comparative Study of GD, LM and SCG of Neural Network for thyroid disease diagnosis”. International Journal of Advanced Research; 2015

[23] Banu GR. “Predicting Thyroid Disease using Linear Discrimination Analysis (LDA) data mining techniques”; 2016

[24] F. S. Gharehchopogh, M. Molany and F. D.Mokri, ”Using Artificial Neural Network In Diagnosis Of Thyroid Disease: A Case Study”, International Journal on Computational Sciences &Applications (IJCSA) Vol.3, No.4, August 2013

[25] Francois, D., Rossi, F., Wertz, V., Verleysen, M. Resampling methods for parameter free and robust feature selection with mutual information. Neurocomputing 70, 1276-1288 (2007).

[26] Diamantidis, N.A., Karlis, D., Giakoumakis, E.A. Unsupervised stratification of cross validation for accuracy estimation. ArtifIntell; 116:1-16 (2000).

[27] Salzberg, S.L. On comparing classifiers: Pitfalls to avoid and a recommended approach. Data Mining and Knowledge Discovery, 1, 317-327 (1997).