International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 2 pp. 787-793 June 2018 www.csjournals.com

Prediction of Thyroid Disease Using Machine Learning Techniques

[1]SunilaGodara, [2]Sanjeev Kumar
[1,2]Department of CSE, Guru Jambheshwar University of Science & Technology, Hisar, India

Abstract: Hypothyroidism or hyperthyroidism is a major disease in India which arises due to malfunctioning of thyroid hormones. Medical industry has enormous quantity of data, but the bulk of this data is not processed. For proper diagnosis data must be processed accurately. For accurate processing intelligent Machine Learning techniques are widely used. In this paper an attempt is made to analyze Logistic regression andSupport Vector Machine (SVM) for multiclass classification of thyroid dataset.Performance of these techniques ison basis of Precision, Recall, F measure, ROC, RMS Error and accuracy. Our analysis shows that logistic regression is more efficientthanSVM for multiclass classification of thyroid dataset.

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,sadfatigue, depression, constipation, feeling cold, weight gain, muscle weakness, dry, thinning hair,slowed heart rate). Hyperthyroidism is a condition when thyroid gland produces too much thyroid hormones[2]. Symptoms of hyperthyroidism are nervousness, restlessness, inability to concentrate ,increased appetite, difficulty sleeping, itching, hair loss, nausea and vomiting. For diagnosis entire medical history and physical tests (free T4, T3Test, Cholesterol test, TSH Test)are required. As these test produces large amount of data and MLcan be used forfinding important features from large amount of data. Due to this specialty of ML can be usedin combination with medical science for the accurate diagnosis of hypo thyroidismdisease[1]. A number of ML techniqueshave been evolved and in order to achieve best accuracy of a model ensembles are widely used[7].

A nonparametric test exposed accurately massive contrasts between FV-PTC and ordinary thyroid using both parameters (p <; 0.05). These preliminary results advised that phantom based quantitative ultrasound imaging using machine learning is valuable for the administration of thyroid tumour [4]. The outcomes demonstrates that the JET model providedexact depictions of thyroid knobs when contrasted with LD and copes with the confinements of the past thyroid depiction approaches [5].Polat considered artificial immune- recognition system (AIRS) for thyroid diagnosis, and found 81% accuracy [6]. Keles proposed an expert system using Neuro-Fuzzy classification method for thyroid diagnosis and found an accuracy of 95.33% [8].Temurtas did thyroid disease diagnosis with the help of Multi Layer Perception (MLP) with Levenberg Marquardt- LM algorithm was done and found accuracy of 93.19%[9]. Wavelet used Support Vector Machine (WSVM) and Generalized Discriminant Analysis (GDA) methods for thyroid diagnosis and got 91.86% classification accuracy[10]. Chen did optimization using particle swarm optimization for thyroid disease, and found accuracy of 97.49% [11].Chen,Hui-Ling proposed an expert system, called Fisher Score Particle Swarm Optimization Support Vector Machines (FS-PSO-SVM) and was evaluated on thyroid disease dataset [12].Binary logistic regression, naïve Bayes classifier,support vector machine (SVM), and radial basis function neural network (RBFNN) were analyzed forthyroid diagnosis[13–14]. Analyzed various ML techniques for medical diagnosis[15].

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 2 pp. 787-793 June 2018 www.csjournals.com

II. Thyroid Detection Models

Under this section we will discuss Logistic Regressionand SVM modelsused to detect Thyroid Disease.

A. Logistic Regression

Logistic regression is a ML technique used to allocate records into discrete set of classes. Linear regression produces continuous number values as output. Linear Regression can predicts the student's test attain on a scale of 0 - 100 (predictions are continuous as range is required). Logistic Regression[1] can be used to predict whether the student is pass or fail. Logistic regression predictions are discrete (only exact values or categories are permissible). Binary logistic regression: this will take two values 0 or 1

$$Y = b_0 + b_1 X + e$$

To map predicted values to probabilities, sigmoid function is used. This function maps any real value into another value between 0 and 1. In machine learning, sigmoid is used to map predictions to probabilities.

$$\sigma^t = \frac{1}{1 + e^{-t}}$$

The output of this is the estimated probability whichtells how confident can predicted value be actual value when X is given as an input.

B. Support Vector Machine

SVM[12] is used for the classification of both linear and non-linear data. This technique is derived from statistical learning theory given by Vipnik in 1992. SVM technique solvesthe problem by finding out the hyper-plane with maximum margin.For nonlinearlyseparable data, it transformsthe training data into a higher dimension space by doing non linear mapping. By transforming it into high dimensional space, it searches for linear optimal separating hyper-plane. This transformation technique into high dimension always helps in searching for an optimal hyper-plane using support vectors and margins[13]. SVM achieved classification by finding optimal MMHand minimizing the classification errors.

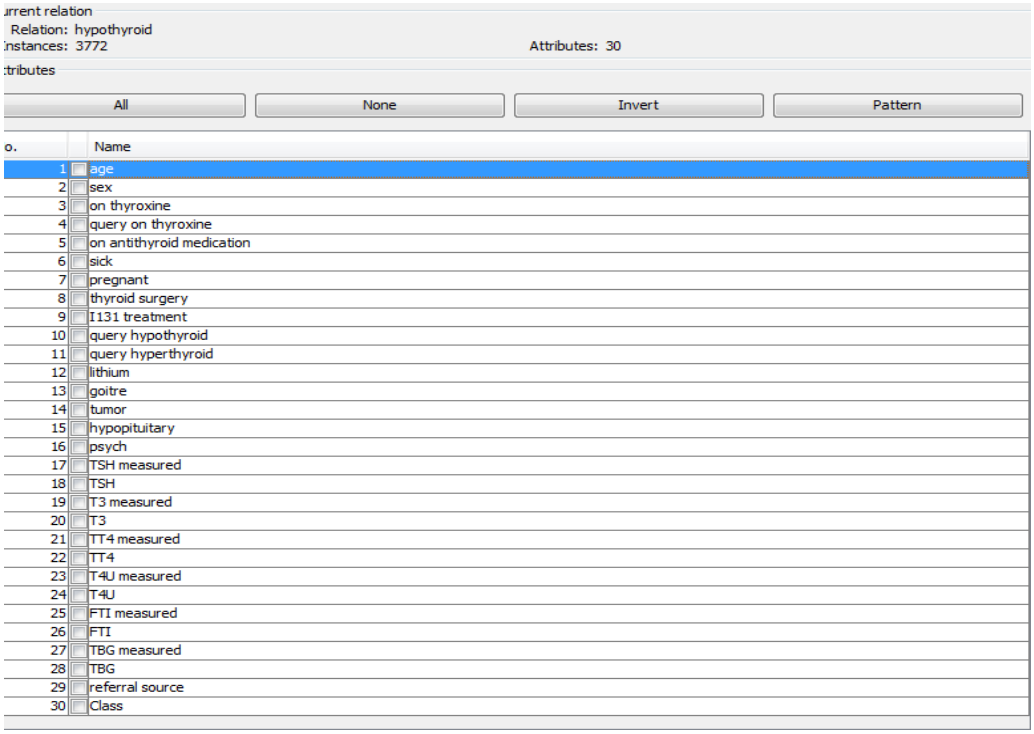

Figure1.Thyroid dataset

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 2 pp. 787-793 June 2018 www.csjournals.com

III. Data Source

To detect Thyroid Disease, dataset wastaken from UCI repository. The Thyroid dataset has 30 attributes and 3772 records.

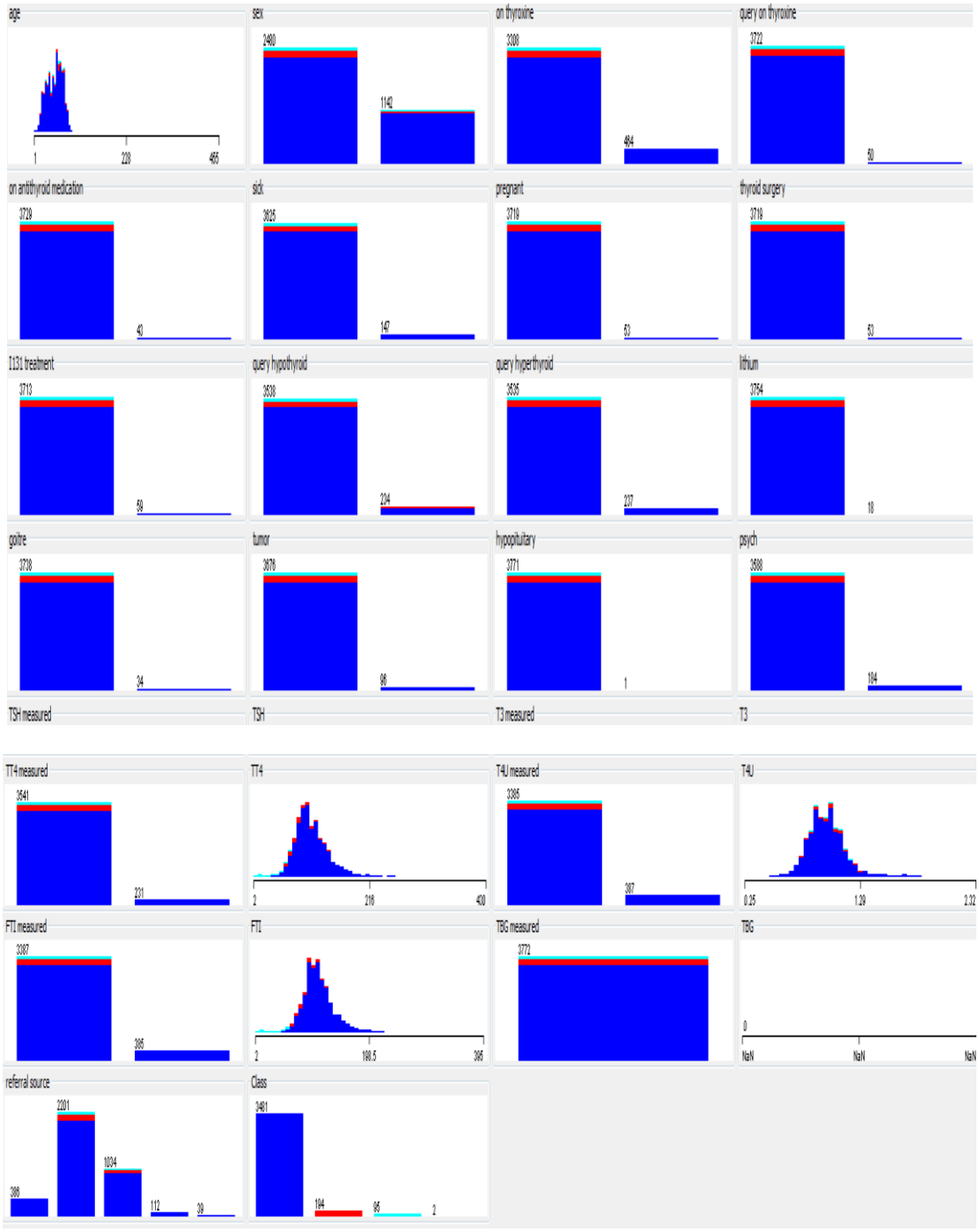

Fig2: Attributes description of Thyroid dataset

In Thyroid dataset,Class is nominal variable having four different values. From above Figure we can find that age, TSH, T3,TT4 ,T4U,FTI,TBG,Referral source are numeric variables.Sex, on thyroxine,

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 2 pp. 787-793 June 2018 www.csjournals.com

query on thyroxine,on antithyroid medication,sick, pregnant,thyroid surgery,I131 treatment, query hypothyroid, query hyperthyroid,lithium,goitre and all the remaining attributes are nominal having two values.

IV. Results

Logistic regression and SVM machine learning techniques are used to analyze Thyroid dataset using Weka version 3.6. Initially dataset had 30 attributes and 3772 records. Logistic regression and Support Vector Machine are compared on basis of Precision,Recall,F measure,ROC and RMS Error.Figure 1. shows confusion matrix obtained by logistic regression. Figure 2. shows results obtained usinglogistic regression..Figure 3 shows confusion matrix obtained by SVM.Figure 4 shows results obtained using SVM.

```
=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 3459    6    9    7 |    a = negative
   74  118    2    0 |    b = compensated_hypothyroid
    5   13   76    1 |    c = primary_hypothyroid
    1    0    1    0 |    d = secondary_hypothyroid
```

Figure 1:Confusion matrix obtained by Logistic Regression

```
Correctly Classified Instances        3653               96.8452 %
Incorrectly Classified Instances       119                3.1548 %
Kappa statistic                          0.7604
Mean absolute error                      0.0256
Root mean squared error                  0.112
Relative absolute error                 35.0862 %
Root relative squared error             58.7932 %
Total Number of Instances             3772

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.994    0.275      0.977     0.994     0.985      0.98     negative
                 0.608    0.005      0.861     0.608     0.713      0.983    compensated_hypothyroid
                 0.8      0.003      0.864     0.8       0.831      0.971    primary_hypothyroid
                 0        0.002      0         0         0          0.76     secondary_hypothyroid
Weighted Avg.    0.968    0.254      0.968     0.968     0.967      0.979
```

Figure 2: Results obtained by Logistic Regression

```
=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 3479    0    2    0 |    a = negative
  193    0    1    0 |    b = compensated_hypothyroid
   43    0   52    0 |    c = primary_hypothyroid
    2    0    0    0 |    d = secondary_hypothyroid
```

Figure 3:Confusion matrix obtained by SVM

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 2 pp. 787-793 June 2018 www.csjournals.com

```
Correctly Classified Instances        3531               93.6108 %
Incorrectly Classified Instances       241                6.3892 %
Kappa statistic                          0.292
Mean absolute error                      0.256
Root mean squared error                  0.3213
Relative absolute error                351.2674 %
Root relative squared error            168.7332 %
Total Number of Instances             3772

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.999     0.818      0.936     0.999     0.967      0.591    negative
                 0         0          0         0         0          0.519    compensated_hypothyroid
                 0.547     0.001      0.945     0.547     0.693      0.86     primary_hypothyroid
                 0         0          0         0         0          0.499    secondary_hypothyroid
Weighted Avg.    0.936     0.755      0.888     0.936     0.91       0.594

=== Confusion Matrix ===
```

Figure 4: Results obtained by SVM

	Precision	Recall	F measure	ROC	RMS Error
Logistic regression	.968	.968	.967	.979	.112
SVM	.888	.936	.91	.594	.3213

Table 1: Comparison of Logistic regression and Support Vector Machine (SVM) on basis of Precision,Recall,F measure,ROC and RMS Error

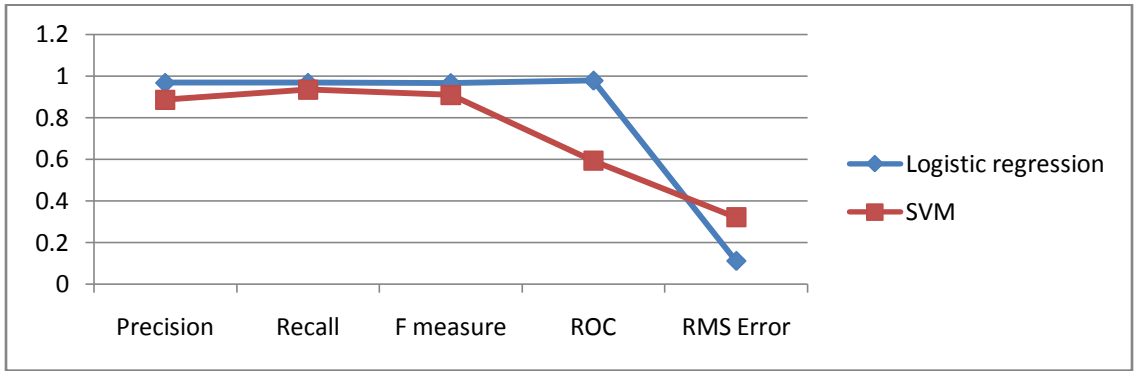

Figure5: Graphical representation ofComparison of Logistic regression and Support Vector Machine (SVM) on basis of Precision,Recall,F measure,ROC and RMS Error.

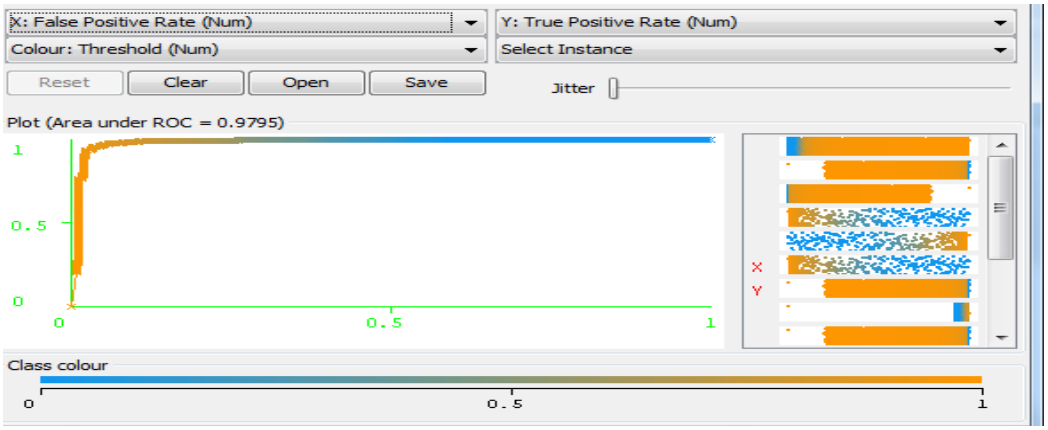

Figure 6: ROC curve of Logistic regression

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 10 • Issue 2 pp. 787-793 June 2018 www.csjournals.com

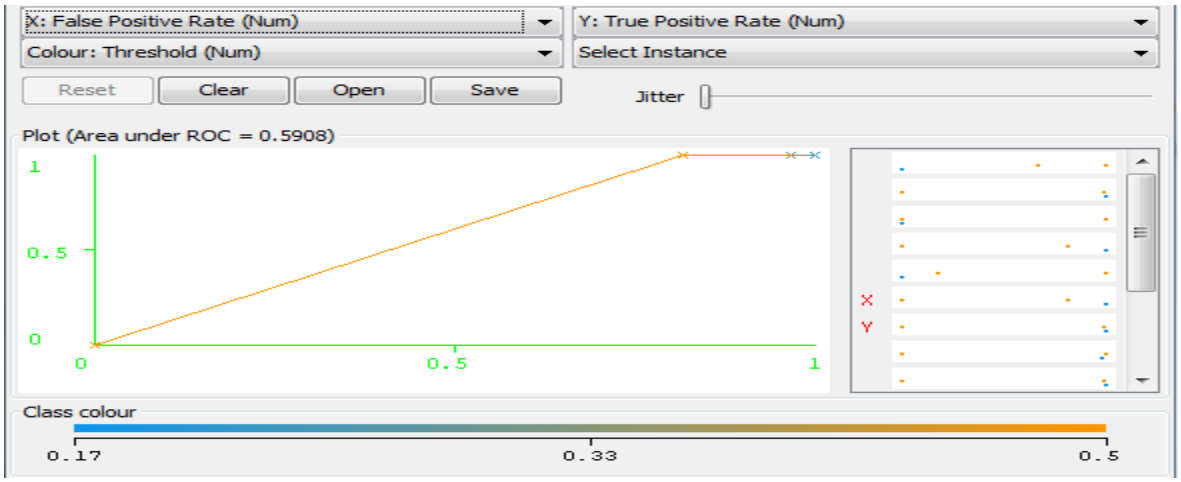

Figure 7: ROC curve of SVM

Accuracy	
Logistic Regression	96.8452
SVM	93.6108

Table 2: Comparison of Logistic regression and Support Vector Machine (SVM) on basis of Accuracy

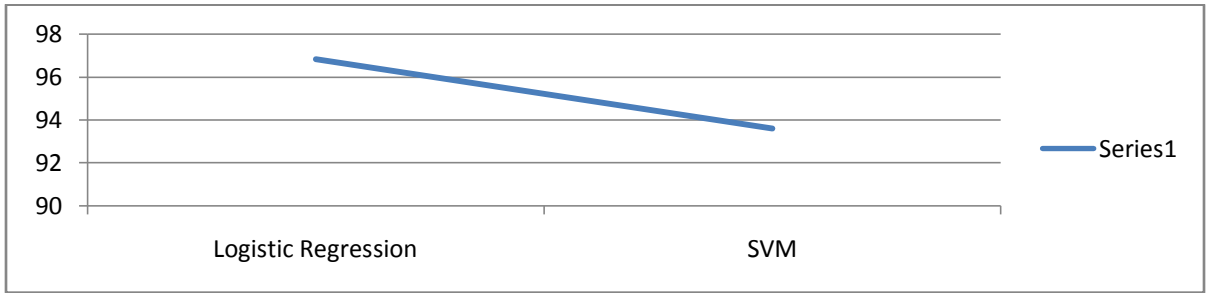

Figure 8: Graphical representation of Comparison of Logistic regression and Support Vector Machine (SVM) on the basis of Accuracy.

Table 1. depicts the comparison between Logistic regression and Support Vector Machine (SVM) on basis of Precision, Recall, F measure, ROC and RMS Error.Figure 5 shows the Graphical representation of Logistic regression and Support Vector Machine (SVM) on basis of Precision,Recall,F measure,ROC and RMS Error. Figure 5 showsthat logistic regression is performed better than Support Vector Machine in allthe parameters like Precision,Recall,F measure,RMS Error.Figure 6 and Figure 7 shows the ROC curve of Logistic regression and Support Vector Machine (SVM). Logistic Regression outperformed Support Vector Machine as shown in ROC curve. Table 2 exhibits the comparison between Logistic regression and Support Vector Machine (SVM) on basis of Accuracy.Figure 8 shows the Graphical representation of Logistic regression and Support Vector Machine (SVM) on basis of accuracy. Figure 8 shows that logistic regression is performed better than Support Vector Machine on the basis of accuracy. No doubt, SVM is more stable technique than Logistic Regression for binary classification. But in case of multiclass classification Logistic Regression outperforms SVM.Hence, it can conclude that SVM performance deteriorates as number of classes increases.

V. Conclusion:

ML techniquescan be used for Thyroid detection. In this paper logistic regression and SVM are usedto predictThyroid.These techniques are compared on the basis of Precision,Recall,F measure,ROC, RMS

Error and accuracy. This paper showed that instead of SVM,logistic regression turns out to be best classifier for Thyroid detection when number of classes increases.

References

1. S. Godara and R. Singh, "Evaluation of Predictive Machine Learning Techniques as Expert Systems in Medical Diagnosis", Indian Journal of Science and Technology, Vol. 910, March 2016.
2. Zhang, Guoqiang Peter, and Victor L. Berardi. "An investigation of neural networks in thyroid function diagnosis." Health Care Management Science Vol. 1, Issue 1, Sep 1998.
3. Sunila, Rishipal Singh and Sanjeev Kumar. "A Novel Weighted Class based Clustering for Medical Diagnostic Interface." Indian Journal of Science and Technology Vol 9, Issue 44, Nov.2016.
4. Lavarello, Roberto J., Billy Ridgway, Sandhya Sarwate, and Michael L. Oelze. "Imaging of follicular variant papillary thyroid carcinoma in a rodent model using spectral-based quantitative ultrasound techniques." In Biomedical Imaging (ISBI), IEEE 10th International Symposium, pp. 732-735, 2013.
5. Savelonas, Michalis A., Dimitris K. Iakovidis, IoannisLegakis, and Dimitris Maroulis. "Active contours guided by echogenicity and texture for delineation of thyroid nodules in ultrasound images." IEEE Transactions on Information Technology in Biomedicine,Vol. 13, Issue 4, July 2009.
6. Polat, Kemal, Seral Şahan, and SalihGüneş. "A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted pre-processing for thyroid disease diagnosis." Expert Systems with Applications, Vol 32, Issue 4,May 2007.
7. Kumari, Milan, and SunilaGodara. "Comparative study of data mining classification methods in cardiovascular disease prediction 1." IJCST, Vol. 2, 2011 ISSN: 2229-4333
8. Keleş, Ali, and AytürkKeleş. "ESTDD: Expert system for thyroid diseases diagnosis." Expert Systems with Applications Vol 34, Issue 1, 2008.
9. Temurtas, Feyzullah. "A comparative study on thyroid disease diagnosis using neural networks." Expert Systems with Applications, Vol 36, Issue 1, 2009.
10. Dogantekin, E., Dogantekin, A., and Avci, D., "An expert system based on generalized discriminant analysis and wavelet support vector machine for diagnosis of thyroid diseases." Expert Syst. Appl. , Vol. 38, Issue 1, 2011.
11. Stegmayer, Georgina, Matias Gerard, and Diego H. Milone. "Data mining over biological datasets: An integrated approach based on computational intelligence." IEEE Computational Intelligence Magazine, Vol 7, Issue. 4 Nov 2012.
12. Chen, Hui-Ling, Bo Yang, Gang Wang, Jie Liu, Yi-Dong Chen, and Da-You Liu. "A three-stage expert system based on support vector machines for thyroid disease diagnosis." Journal of medical systems, Vol 36, Issue 3, 2012.
13. Yoo, Young Jin, EunJu Ha, Yoon Joo Cho, Hye Lin Kim, Miran Han, and So Young Kang. "Computer-Aided Diagnosis of Thyroid Nodules via Ultrasonography: Initial Clinical Experience." Korean Journal of Radiology, Vol 19, Issue 4 2018.
14. Chen, Shao-Jer, Chuan-Yu Chang, Ku-Yaw Chang, Jeh-EnTzeng, Yen-Ting Chen, Chih-Wen Lin, Wen-Ching Hsu, and Chang-Kuo Wei. "Classification of the thyroid nodules based on characteristic sonographic textural feature and correlated histopathology using hierarchical support vector machines." Ultrasound in medicine & biology, Vol 36, Issue 12, 2010.