A Study on Label TSH, T3, T4U, TT4, FTI in Hyperthyroidism and Hypothyroidism using Machine Learning Techniques

Conference Paper · July 2019

DOI: 10.1109/ICCES45898.2019.9002284

6 authors, including:

Afzal Shahid
National Institute of Technology Patna

A Study on Label TSH, T3, T4U, TT4, FTI in Hyperthyroidism and Hypothyroidism using Machine Learning Techniques

Afzal Hussain Shahid[1], Maheshwari Prasad Singh[1], Rahul Kumar Raj[2], Rashmi Suman[2], Drakhshan Jawaid[2], Muqtadir Alam[2]

[1]Department of Computer Science and Engineering, National Institute of Technology Patna, Patna, India
({afzal.cse15, mps}@nitp.ac.in)

[2]Department of Computer Science and Engineering, Muzaffarpur Institute of Technology Muzaffarpur, India
({rahulrajab525, sumanrashmi2017, alammit2015, drakhshan933}@gmail.com)

Abstract— **The thyroid hormone is produced by thyroid gland. This hormone regulates the body's metabolism. Hyperthyroidism and hypothyroidism are the two abnormalities which is caused by the release of too much or too little thyroid hormones respectively. In this study, Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbours (K-NN) classifiers are compared to assess the efficiency of these classifiers in Thyroid disease diagnoses using the thyroid disease dataset that is taken from UCI machine learning repository. The overall classification accuracy of the RF, SVM, and K-NN are 98.50%, 97.02%, and 95.81% respectively. The result shows that the RF classifier performance is better than SVM and K-NN for the diagnosis of thyroid disease using UCI dataset.**

Keywords— Thyroid disease diagnosis, Computer-aided diagnosis, Machine learning classifiers.

I. INTRODUCTION

The thyroid gland is one of the endocrine glands that produce thyroid hormone to regulate the body's metabolism [1]. The Levothyroxine (T4) and Triiodothyronine (T3) hormones are produced by thyroid gland. These hormones play a major role in converting nutrients into energy and thereby helps in regulating body temperature, heart rate, and even brain function. Hyperthyroidism occurs when the hormones produced by the gland is more than the requirement of the body. However, hypothyroidism occurs when the hormones produced by the gland is less than the requirement of the body [2]. Severe hyperthyroidism and end-stage hypothyroidism may cause deaths [3] [2]. During pregnancy, it can lead to health problems for the mother and baby [4].

About 70% risk of developing the thyroid disease is found to be associated with genetic background. Also, the contribution of environmental exposures such as perchlorate in rocket fuel, polychlorinated bi-phenols, and radiations (from nuclear fallout and medical radiation) in the disease risk is significant. Most of the environmental exposures reduce thyroid hormone levels [5]. Recently, the thyroid disease risk is found to be linked with systemic lupus erythematosus, an autoimmune disease [6]. There is some evidence of the correlation between thyroid disease and the low level of vitamin D [7].

The amount of T3 and T4 hormone in the blood is controlled by another hormone known as Thyroid Stimulating Hormone (TSH). Low T4 results in more production of TSH that initiates thyroid gland to produce more T4 & vice versa [8].

The different thyroid function tests such as TSH, T3, T4U, Total T4, and Free T4 Index (FTI) are done to assess the various functions related to the thyroid gland. A normal TSH and normal T4 is indicative of normal functioning of the thyroid gland, a low TSH and elevated T4 indicates hyperthyroidism, a low TSH and low T4 indicates secondary hypothyroidism, and a high TSH and low T4 indicates primary hypothyroidism. T3 tests are used for determining the severity of the hyperthyroidism or to diagnose the hyperthyroidism [9].

The classification algorithm is based on supervised learning methodology where the labeled data is used for learning. These algorithms require very large labeled training sets which consist of many features or attributes.

Finding meaningful information from the thyroid datasets of the patients with regard to the thyroid disease diagnosis is important. This paper compares the performance of three classification algorithms, namely, RF, SVM, and K-NN for the diagnosis of thyroid disease by analyzing labels (TSH, TT4, T3, FT4, and T4U) in hyperthyroid, hypothyroid, and normal subjects.

The rest of the paper is organized as follows. Section II briefly describes the related work. Section III presents the methodology that includes the description of the dataset used, a brief introduction about ML classifiers (RF, SVM, K-NN), and the performance metrics used in the analysis. Section IV presents the results and discussions. Section V concludes this paper.

II. RELATED WORK

Researchers have used various machine learning algorithm for the diagnosis of thyroid diseases. For instance, authors of [10] use Wavelet Support Vector Machine and Generalized Discriminant Analysis to develop an expert system for thyroid disease diagnosis. They achieved the classification accuracy of about 91.86%. Polat et al. [2] hybridize the artificial immune recognition system with fuzzy weighted pre-processing to diagnose the thyroid disease. The classification accuracy of their model was 85%. Feyzullah Temurtas [3] suggests the diagnosis of thyroid disease using Neural Networks. They obtained the accuracy of 93.19%, 94.81%, and 90.05% using multilayer, probabilistic, and learning vector quantization neural networks respectively. Gharehchopogh et al. [11] propose a Multi-layer Perceptron (MLP) which is trained via back propagation learning algorithm for the classification of thyroid disease. The classification accuracy of 96.8% was achieved.

III. METHODOLOGY

A. Dataset description

The dataset for this study is acquired from the UCI machine learning repository. The dataset consists of 7200 instances, 21 attributes of which 15 attributes are categorical and six attributes are numerical. All the instances in the dataset are labelled to one of the three classes: hyperthyroidism, hypothyroidism, and normal subject.

Table 1. Thyroid disease dataset description.

Number of attributes & instances	Values
Total number of attributes	22 including one label set
Attributes with categorical values	15 with two possible values 0/1
Attributes with numerical values	6
Total number of instances	7200
Number of attributes used for analysis	5

The number of instances belongs to hyperthyroidism, hypothyroidism, and normal subject classes are 166, 368, and 6666 respectively. The distribution of instances among three classes shows that the dataset is highly imbalanced. Further for the analysis purpose, the whole dataset was split into a training dataset and test dataset in the ratio of 80/20. Therefore, there are 5760 instances in the training dataset and 1440 instances in the test dataset. Out of the 21 attributes, only 5 attributes are used for the analysis. These attributes represent the various thyroid function tests that indicate the level of thyroid hormones in the blood and are indicative of thyroid disease. These attributes are as follows [12]:

(i) TSH that ranges between 0.4 to 5.0 milli-International Unit per liter.
(ii) Total T4 (TT4) ranges from 4.6 to 12 micrograms per deciliter of blood.
(iii) T3 that ranges between 100 to 200 nanogram per deciliter of blood.
(iv) Free T4 Index (FTI) or FT4 that ranges between 0.7-1.9 nanogram per decilitre of blood.
(v) T4 Uptake (T4U).

B. Support Vector Machine

SVM constructs the hyperplanes in a multidimensional space for the classification task. This space separates classes of different class labels. It is able to deal with non-linear and high-dimensional data. With RBF kernel, the performance is further improved in case of non-linear and high dimensional data. The drawbacks of SVM include: (i) selection of kernel function is difficult as the performance varies with different kernel functions and datasets, and (ii) time-consuming training phase incurs high computational cost [13]. The extreme data points in each class represent the support vectors while the decision boundary between them is the hyperplane which is drawn in such a way that the distance between the support vector and the plane should be maximum.
In this study, the SVM model is trained by using the training samples and the model is tuned by tuning the various parameters to get the optimum result. The parameters used for tuning are Kernel function, C parameter, and gamma coefficients. Kernel function does the mapping of data to different space. The various kernel functions are Linear, rbf, poly, sigmoid, precomputed or a callable. In this study, the SVM model is tuned by using rbf and sigmoid kernel function. The parameter C is the regularization parameter that has the ability to generalize the SVM to perform well on the unseen data. It controls the tradeoff between the achieving a low training error and a low testing error. A large value of C corresponds to a small margin whereas a small value of C corresponds to the large margin between support vectors and hyperplane. In this study the SVM model is tuned at various values of C. Gamma is the kernel coefficient for rbf, poly, sigmoid kernel function which adds complexity in the classification.

C. Random Forest

Random forest (RF) builds multiple decision trees and combine them toger to obtain a more robust, accurate and stable prediction. It adds additional randomness to the model while growing the trees. This algorithm searches for the best feature among a random subset of features for splitting a node.

In RF classification approach, voting strategy is used where each tree votes for a particular class. The rows or instances are then classified into the class for which maximum number of trees have voted. Due to more robustness, the performance of RF it better than AdaBoost, which is also an ensemble learning technique. The performance of RF is better than DT because, when there is noise in the attribute, trees become deeper and hence overfitting problem occurs. However, RF is less influenced by the noise in the attributes as different sample of the training data is used in the process of averaging the deep decision trees [13]. In this study, the RF model is tuned by varying the n_estimators parameter which represents the number of trees in the forest.

D. K-Nearest Neighbor

K-NN classifies a data point based on how its neighbors are classified. It computes the distance between the new instance and previously known instance to classify the new instance by using the voting strategy. The information about previous instances is given in the training set. It is simplest among all ML classifiers [13]. In this study, the K-NN model is tuned by varying the value of K and compared with each other to get the best classification performance.

E. Performance Metrics

Following performance metrics are used to analyze the performance of different ML classifiers.

Precision: It is used to measures the classifier exactness. It is also known as positive predictive value (PPV).

Recall: It measures the classifier completeness. It is also known as sensitivity or hit rate, or true positive rate (TPR).

Specificity: It measures the proportion of actual negatives that are correctly identified and also called a true negative rate (TNR).

Accuracy: It measures the fraction of predictions our model got right. It can be calculated as the total number of correct predictions divided by the total number of test cases.

F1-Score: It measures the balance between precision and recall.

Class weighted accuracy: It can be calculated as the ratio of the sum of each class size multiplied by its individual accuracy to the total number of test cases.

IV. Results and Discussion

The aim of this paper is to assess the classification efficiency of different ML algorithms on the thyroid disease dataset acquired from UCI. Table 2 shows the classification performance of different ML algorithms in terms of precision, recall, specificity, accuracy, and F1-score for each class. All the models were tuned to enhance the performance. The SVM gives the best result with the *RBF* kernel function and when the values of the parameters C and gamma coefficient are 100 and 1024 respectively. The RF gives the best result when the value of the parameter n_estimators is 32 and K-NN gives the best result when the value of the parameter n_neighbors is set to 1.

It can be observed from Table 2 that for all the classes the RF model gives the best performance in terms of all the performance metrics. It was seen that the performance of the SVM model was improving on increasing the value of the parameter C and best performance is achieved when the value of C is 100 but further increase in the C value results in the performance deterioration. Because the optimal value of C parameters should be large when the feature values in the dataset is small. However, for large value of C, the misclassification rate increases which results in poor performance. Similarly, for RF, the performance increases by increasing the value of the n_estimators as for the higher value of n_estimators, the performance of RF increases. When the value of n_estimators is 32, the best performance is achieved. The further increase in the value of n_estimators results in no change in the performance. However, for the K-NN model, the performance is best at k=1 and starts decreasing by further increasing the value of K. The overall class weighted performance measures for all the classification model is presented in Table 3.

Table 2: Performance of different ML algorithms.

Classes	**Performance metrics**	**Classification models**		
		SVM (%)	**RF (%)**	**K-NN (%)**
Hyperthyroidism	Precision	91.66	92.68	84.84
	Recall	84.61	97.43	71.79
	Specificity	99.78	99.78	99.62
	Accuracy	99.35	99.71	98.84
	F1-Score	88	95	77.77
Hypothyroidism	Precision	76.8	86.51	69.56
	Recall	75.90	92.77	57.83
	Specificity	98.59	99.11	98.44
	Accuracy	97.27	98.74	63.15
	F1-Score	76.36	89.53	96.08
Normal subjects	Precision	98.18	99.46	97.01
	Recall	98.48	98.86	98.48
	Specificity	80	94.26	65.51
	Accuracy	96.94	98.47	97.74
	F1-Score	96.94	99.16	95.81

Table 3: Overall class weighted performance for the different classification model.

Overall Class Weighted Performance Measures	**Classification models**		
	SVM (%)	**RF (%)**	**K-NN (%)**
Precision	96.77	**98.54**	95.26
Recall	96.80	**98.57**	95.49
Specificity	81.60	**94.69**	70.67
Accuracy	97.02	**98.50**	95.81
F1-Score	96.78	**98.52**	95.35

Figure 1 shows the overall class weighted average performance of SVM, RF, and K-NN with respect to the different performance parameters. For all the performance parameters (precision, recall, specificity, accuracy, and F1-

score), the best performance is achieved by RF followed by SVM and K-NN.

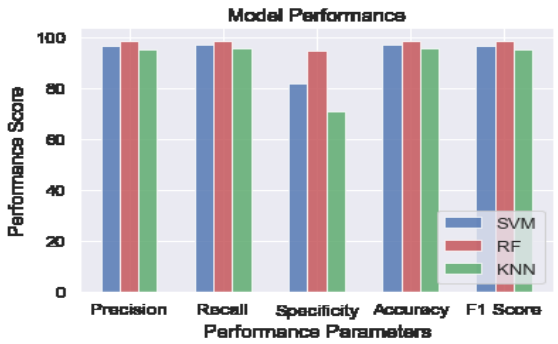

Fig. 1. Overall performance measure comparison for all the classification model

V. CONCLUSION

The paper presented a comparison of three different ML techniques namely, SVM, RF, and K-NN applied for the classification of the thyroid disease into hyperthyroidism, hypothyroidism, and normal subjects. The dataset for this study was acquired from the UCI machine learning repository. To evaluate the performance of different ML algorithms, various performance metrics such as precision, recall, specificity, accuracy, and F1-score were used. Although all the models may be used for the classification, the RF classifier outperformed SVM and K-NN. The maximum weighted average accuracy achieved by RF was 98.50 %. However, the accuracies achieved by SVM, and K-NN are 97.02%, and 95.81% respectively. RF also gives the best performance on all the performance metrics in comparison to SVM and KNN. In future, it would be interesting to assess the classification performance of metaheuristic algorithms on this dataset.

ACKNOWLEDGMENT

The authors would like to acknowledge the Ministry of Electronics & Information Technology (MeitY), Government of India for supporting the financial assistant during research work through “Visvesvaraya Ph. D. Scheme for Electronics & IT”.

REFERENCES

[1] L. Ozyilmaz and T. Yildirim, "Diagnosis of thyroid disease using artificial neural network methods," in *Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP'02.*, 2002, vol. 4, pp. 2033-2036: IEEE.

[2] K. Polat, S. Şahan, and S. Güneş, "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, no. 4, pp. 1141-1147, 2007.

[3] F. Temurtas, "A comparative study on thyroid disease diagnosis using neural networks," *Expert Systems with Applications,* vol. 36, no. 1, pp. 944-949, 2009.

[4] B. M. Casey and K. J. Leveno, "Thyroid disease in pregnancy," *Obstetrics & Gynecology,* vol. 108, no. 5, pp. 1283-1292, 2006.

[5] G. A. Brent, "Environmental exposures and autoimmune thyroid disease," *Thyroid,* vol. 20, no. 7, pp. 755-761, 2010.

[6] J.-S. Yun *et al.*, "Increased risk of thyroid diseases in patients with systemic lupus erythematosus: A nationwide population-based Study in Korea," *PloS one,* vol. 12, no. 6, p. e0179088, 2017.

[7] D. Kim, "The role of vitamin D in thyroid diseases," *International journal of molecular sciences,* vol. 18, no. 9, p. 1949, 2017.

[8] D. A. Fisher, "Physiological variations in thyroid hormones: physiological and pathophysiological considerations," *Clinical Chemistry,* vol. 42, no. 1, pp. 135-139, 1996.

[9] S. P. Fitzgerald and N. G. Bean, "The relationship between population T4/TSH set point data and T4/TSH physiology," *Journal of thyroid research,* vol. 2016, 2016.

[10] E. Dogantekin, A. Dogantekin, and D. Avci, "An expert system based on Generalized Discriminant Analysis and Wavelet Support Vector Machine for diagnosis of thyroid diseases," *Expert Systems with Applications,* vol. 38, no. 1, pp. 146-150, 2011.

[11] 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,* vol. 3, pp. 49-61, 2013.

[12] A. Keleş and A. Keleş, "ESTDD: Expert system for thyroid diseases diagnosis," *Expert Systems with Applications,* vol. 34, no. 1, pp. 242-246, 2008.

[13] A. H. Shahid and M. P. Singh, "Computational intelligence techniques for medical diagnosis and prognosis: Problems and current developments," *Biocybernetics and Biomedical Engineering,* 2019.