**Classification Algorithms for Diagnosis of Thyroid Disease**

**Abstract:** Thyroid disease is worldwide spread disease. Even in India about 49 million of people are suffering from thyroid disease due to dysfunction of thyroid gland. From the overall perspective, classifying thyroid plays an important role in identifying the type of thyroid disease patient suffers from. Thyroid gland is the most important gland which secretes thyroid hormones to maintain the body’s temperature, heart rate as well as metabolic rate. Several disease due to thyroid disorders occurs like thyroiditis. But our study generally focuses on the hyperthyroidism and hypothyroidism. Hypothyroidism - it is the state of insufficient or too little production of hormones. Hyperthyroidism - when glands produces excessive amount of thyroid hormones. The condition for diagnosis of the thyroid disease are closely attached; they have several important issues that could affect diagnosis and treatment of a person. The dataset taken from the uci repository which undergoes for pre-processing. The pre-processing outcome of data is uncertain in its characteristic.

**Keywords:** SVM, K-NN, complex decision trees, thyroid disease, confusion matrix.

1. **Introduction**

Thyroid gland is an endocrine gland found in the lower part of human neck, which helps in secretion of thyroid hormones, and maintaining and balancing the body’s metabolism [1]. Generally two types of hormones are produced by thyroid glands, namely levothyroxine (T3) and triiodothyroxine (T4). The functionalities of these two hormones are that it helps in production of well-balanced amount of proteins, regulating the body’s temperature, and maintaining overall production of energy. Thyroid disease can be divided normally into two types, these are hypothyroidism and hyperthyroidism. Hypothyroidism - it is the state of insufficient or too little production of hormones. Hyperthyroidism - when glands produces excessive amount of thyroid hormones, commonly causing Grave’s disease [2].

Thyroid gland looks similar to the shape of a butterfly which have two wings being represented by the left and right thyroid lobed which are wrapped around the trachea. The main function of the thyroid is to make thyroid hormones [3]. This hormone affects nearly all tissues of the body where it increases activities like cellular activity as well as metabolism activity. The thyroid data are obtained from the measurements of the thyroid gland. The thyroid makes two types of active thyroid hormones i.e., levothyroxine (also called T4) and triiodothyronine (also called T3). The functionalities of these two kinds of hormones are that its helps in production of balanced amount of proteins, regulating the temperature of body, and maintaining overall production of energy. The thyroid gland suffers from many unique problems, most of them are very individual while most of them are very usually occurred. Production of less or little amount of thyroid hormone results in hypothyroidism and secretion of too much or excessive amount of thyroid hormones can cause hyperthyroidism. Hypothyroidism, has many effects. Some of them are before surgery of thyroid, openess to radiation ionizing, chronically inflammation of the thyroid glands or auto-immune thyroid, deficiency of iodine, also less secretion of enzymes to make thyroid hormones. Hyperthyroidism, or called as an overactive thyroid, can also be caused due to local physical condition of the thyroid, various methods for medications, and lack of control of thyroid hormones secretion. One of the most common cause occurred due to hyperthyroidism is graves disease. Graves disease occurs when the body makes proteins that continuously tells the thyroid to make even more secretion of thyroid hormone. The issue related to thyroid disease should never be underestimated by thyroid patient because it may cause disease like thyroid storm (type of critical hyperthyroidism) and myxedema (the last stage of untreated hypothyroidism) which may result in death [4].

Recent studies have shown that women are 5 to 8 times more prone to thyroid disease in comparison from men. Hypothyroidism can even be associated with pregnancy in women as well. For correct diagnosis of thyroid disease, interpretation of thyroid data must be carefully observed beside the clinical examination because even a minute fluctuation in data can cause severe problems [5].

The study aims to diagnose thyroid disease’s using several classifiers mechanism. We found the classifiers like SVM, K-NN, Decision tree on which the dataset of 215 samples are given as input for classification under these classifiers to train the dataset and check behaviour. The given classifiers SVM method provides the accuracy of 96.30% and also based upon this diagnosis it will open the way for various ill disorder diagnosis for future clinically examine data and increase the chance of progress.

1. **Literature Survey**

**2.1 Thyroid Dataset**

In order to perform the study reported in this paper, we need to have thyroid dataset which is taken from the UCI repository (last accessed: 25 September 2016)). The actual reason of using this dataset is because it is one of the most commonly used dataset for classification systems that we need to compare in this study for thyroid diagnosis problem. The dataset which consists of the thyroid disease consists of three classes namely- normal, hyperthyroidism, hypothyroidism with 215 samples [6]. The class distribution is as shown in given table:

|  |  |  |  |
| --- | --- | --- | --- |
| Table 1: Class distribution of the thyroid dataset | | | |
| **Index** | **Class name** | **Class size** | **Class distribution (%)** |
| C1 | Normal | 150 | 69.77 |
| C2 | Hyper | 35 | 16.28 |
| C3 | Hypo | 30 | 13.95 |

**2.2 Previous Work**

For accurate diagnosis several artificial intelligence methods have been used in last two decades like Multi Layer Perception with Back-Propagation method (MLP) and Radial Basis Function (RBF), the obtained classification with their accuracies are 88.3% and 81.69% respectively. In 1997, MLP, Learning Vector Quantization (LVQ), RBF and Probabilistic Potential Function Neural Network (PPFNN) classifier had been employed and the corresponding accuracy were obtained as 36.74%, 81.86%, 72.09% and 78.14% respectively [7].

In 2004, seven different methods consisting of Linear Discriminant Analysis (LDA), with its satisfying learning parameters, which are C 4.5 with parameter c equal to 5, C4.5 with parameter c equal to 95, DIMLP, MLP with double hidden layers and default learning parameters and SIM to test the outcomes on the thyroid disease, and due to this the accuracies rises to 81.34%, 93.26%, 92.81%, 92.94%, 96.24%, 94.86% and 96.86% respectively [8].

In 2006, Polat studied on the artificial immune- recognition system (called AIRS) for the diagnosis of the thyroid disease, and an accrate result of 81% was found in terms of its accuracy [9]. In 2008, Keles proposed an expert system which was majorly based on Neuro-Fuzzy classification method for thyroid disease diagnosis, and formulated an accuracy of 95.33% [10].

In 2009, Temurtas performed the thyroid disease diagnosis with the help of popular technique of that time called Multi Layer Perception (MLP) with Levenberg Marquardt- LM algorithm, and the corresponding accuracy was evaluated to be 93.19% [11].

In 2011, Wavelet Support Vector Machine (WSVM) and Generalized Discriminant Analysis (GDA) methods for diagnosis of thyroid disease were presented, and obtained 91.86% classification accuracy [12]. In 2011, Chen provided a study on a new optimization method known as particle swarm optimization for thyroid disease, and the average accuracy of 97.49% was founded [13].

An innovative unified computation intelligence approach for mining of biological dataset which involves a neural network and progressive computation was proposed by Stegmayer in year 2012. The author’s method uses latest training algorithms with the help of which a priori knowledge of biological information can be obtained [14].

In 2012, Yeh, Wei-Chang restrained Simplified version of Swarm Optimization (SSO) for mining diagnosis of thyroid dataset collected from UCI repository. Close Interval Encoding was added to represent the ruling- structure more efficiently, and the Orthogonal Array Test was added to powerfully prune all rules to avoid over-fitting the training dataset [15].

In same year 2012, Chen, Hui-Ling proposed an expert system, called Fisher Score Particle Swarm Optimization Support Vector Machines (FS-PSO-SVM) has been comprehensively and very well evaluated in opposition of the thyroid disease dataset, which is commonly used among all authors who use artificial learning methods for diagnosis of thyroid disease [16].

In the above results we have seen that mostly all the results were based on neural networks (NN) and Support Vector Machine (SVM), that have achieved significant results on diagnosis of thyroid disease, but the major disadvantage of these techniques are that due to their black-box nature they are hard to understand, how decisions are made by them [17,18].

In this paper we are going to adopt a method of classification using several algorithms like Support Vector Machine (SVM), Decision Complex Tree, Ensemble bagged tree, K-Nearest Neighbour (KNN). In these techniques we are going to generalise the concepts of sensitivity and specificity, which measures the accuracy of the algorithms used.

As we seen above, thyroid disease is a study of endocrinology, which seems to be an integral element in science of medical applications. Thyroid is one of the most common diseases which is frequently not well interpreted and not well-diagnosed [19]. The Proper analyse of the thyroid data in addition to clinical test and alternate check-up is an important issue in the diagnosis of thyroid disease. Various new methods, namely Linear Support Vector (SVM) machine, Decision complex tree, classifiers, Fine K-Nearest Neighbour (FKNN) etc., are used to train the dataset of each patient. The used data in this study is the thyroid dataset which is taken from the UCI machine learning repository. The reason for using this sort of dataset from UCI-repository is that because it is very often used by the other classification technique that we normally use to in comparing our system with diagnosis of thyroid. We found maximum accuracy in Support Vector Machine (SVM) method for this data set in which a more reliable result is found (96.3% accuracy) by using confusion matrix method. If it is compared with classification results of other methods, we found that SVM algorithm gives highest accuracy as compared to Fine-KNN (95.3%), Complex decision tree (91.2%) and Bagged tree (95.3%) of accuracies.

1. **Proposed Work**

The methodologies that we are going to use on dataset for the diagnosis of thyroid disease are Support Vector Machine (Linear SVM, Quadratic SVM, and Cubical SVM), K-Nearest Neighbour (Fine KNN, Medium KNN, Cubic KNN, and Weighted KNN) and Complex Decision Tree. Using these algorithms, we determine the True Positive Rate (TPR) and True Negative Rate(TNR). This gives the maximum accuracy of the algorithms considering the values of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) that we obtain from confusion matrix.

**3.1 Terminologies and Formulae:**

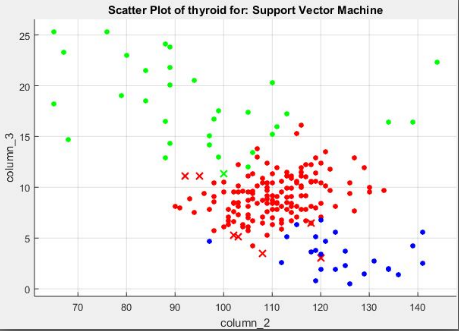
* True positive = correctly identified
* False positive = incorrectly identified
* True negative = correctly rejected
* False negative = incorrectly rejected
* Sensitivity or TPR = TP/(TP+FN)
* Specificity or TNR = TN/(TN+FP)
* Fall Out (also called False Positive Rate) (FPR) = FP/(FP+TN)
* Precision(Positive Predicted Value) = TP/(TP+FP)
* Recall = TP/TP+FN
* Accuracy = (TP+TN)/(TP+TN+FP+FN)

These values are derived from confusion matrix, which is 3x3 matrix with 3 classes normal, hyper and hypo horizontally and vertically where class 1 represents normal, class 2 represents hyper and class 3 represents hypo. The horizontal classes are considered as predicted classes and vertical classes are actual classes. From each cell of confusion matrix we can identify parameters like FP, FN, TP and TN that helps in determining the efficiency of an algorithm to correctly diagnose the thyroid disease.

**3.2 Support Vector Machine**

SVMs are supervised learning technique that is used for classification and regression analysis. It is used to classify the classes based on the model that have group of categories, making it non-probabilistic binary linear classifier. When the data are not labeled, supervised learning is not possible, so unsupervised learning techniques are adopted. For this we use clustering algorithm in order to improve the results of SVMs, which is known as support vector clustering.

|  |  |
| --- | --- |
| Algorithm for SVM: | |
| Step1: | Select candidate = {closest pair of opposite classes} |
| Step2: | while there are violating points do |
| Step3: | find a violator |
| Step4: | candidate = candidate U violator |
| Step5: | if any z < 0 due to addition of c to s then |
| Step6: | candidate = candidate/p |
| Step7: | repeat till all such points are pruned |
| Step8: | end of if condition |
| Step9: | end of while condition |



In the above scatter plot for SVM the various symbols are having some meaning, which can be represented as:

Red dot: shows that class 1 i.e. normal is correctly identified.

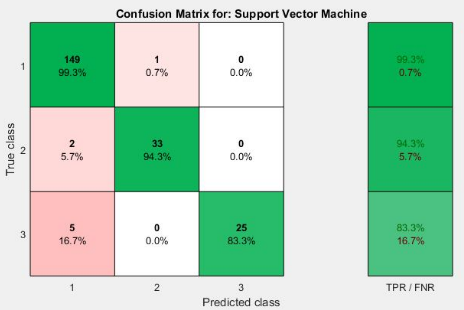
Green dot: shows that class 2 i.e. hyper is correctly identified.

Blue dot: shows that class 3 i.e. hypo is correctly identified.

Red cross: shows that class 1 i.e. normal is incorrectly identified.

Green cross: shows that class 2 i.e. hyper is incorrectly identified.

Blue cross: shows that class 3 i.e. hypo is incorrectly identified.



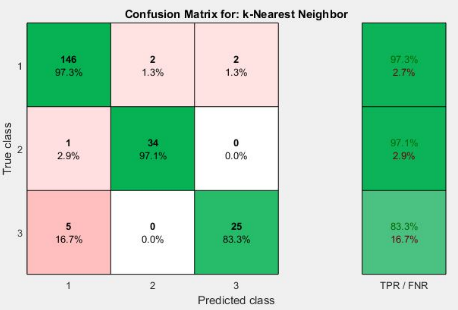
The confusion matrix shows that 149 samples are identified as TP, 58 samples are identified as TN, 7 samples are identified as FP whereas only 1 sample occurs to be FN.

**3.3 K-Nearest Neighbour (KNN)**

It is also known by the name of non-parametric method which is used for classification and analysis of regression of algorithms. In given situations, the input consist of k-training samples in feature-space and output depends on if it is used for classification or regression analysis. K-NN is a type of instance learning or lazy learning. It gives its output based on the dataset that belongs to the same class and are clustered in that specific class.

|  |  |
| --- | --- |
| Algorithm for First K-Nearest Neighbour | |
| Step1: | for each attribute j and each sample k. |
| Step2: | Compute the distance of all remaining attribute. |
| Step3: | values of the attribute j and i: |
| Step4: | for n=1 to N-1 |
| Step5: | Distance (n) = distance (attrbute j, k; attribute j, n)  End for |
| Step6: | find nearest - k attribute values |
| Step7: | find the average values of nearest points which are i and k |
| Step8: | change the value of attribute i and j equal to average of i and j |

|  |  |
| --- | --- |
| Step9: | Next of i |
| Step10: | Next of j |

In this confusion matrix the calculated samples for TP, TN, FP and FN are 146, 59, 6 and 4 repectively.

**3.4 Complex Decision Tree**

It is a tool for making a decision which uses a tree in the form of graph or type of decision and their possible effects. Its structure is similar to flow chart in which all of the internal nodes represents a test on the given attribute and each branch represents the outcome of the resulting test and each leaf node mainly represents a class. The paths from root to leaf represents various classification rules that we have used. In the given confusion matrix for decision tree, the evaluated samples for TP, TN, FP and FN are 144, 52, 13 and 5 respectively.

**3.5 Performance Evaluation**

Using these three algorithms and after evaluating the results, its clearly visible that SVM shows the maximum accuracy with 96.30% as compared to the other algorithms like KNN and decision tree Classifiers with their respective accuracy of 95.34% and 91.16%. These accuracies are measured by considering certain values from confusion matrix like TP, TN, FP and FN as shown above.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 2: Performance parameters | | | |
|  | **SVM** | **KNN** | **Complex Decision tree** |
| Sensitivity | 99.33333 | 97.33333 | 96 |
| Specificity | 89.23077 | 90.76923 | 80 |
| Accuracy | 96.27907 | 95.34884 | 91.16279 |
| Precision | 0.955128 | 0.960526 | 0.917197 |
| Recall | 0.993333 | 0.973333 | 0.96 |
| Fallout | 0.107692 | 0.092308 | 0.2 |

1. **Result**

The classification accuracies obtained from these algorithms and other studies that we have seen in section of literature survey are presented in table 3. The used algorithm especially SVM gives an accuracy of 96.30% which shows better results as compared to previous work of authors like Ozyilmaz and Yildirim in 2002, Polat in 2007, Keles in 2008 and many other popular work of that decade, which is illustrated in given table 3. It differs from other algorithms as SVM focuses only those points which are very much difficult to tell apart, whereas other classifiers pay attention towards all other points in the plane. The other reason for SVM’s better accuracy is that when we get new samples or points, we already made a line that separates the different classes as far away from each other as possible, so the chances are very much minimised that one class will spill over another.

Table 3: Results comparison

|  |  |  |
| --- | --- | --- |
| Study | Method | Accuracy (%) |
| Serpen | MLP  LVQ  RBF  PPFNN | 36.74  81.86  72.09  78.14 |
| Ozyilmaz and  Yildirim | MLP with back-propagation  MLP with fast back-propagation  RBF  CSFNN | 86.33  89.80  79.08  91.14 |
| Pasi | LDA  C 4.5 -1  C 4.5 -2  C 4.5 -3  MLP  DIMLP | 81.34  93.26  92.81  92.94  96.24  94.86 |
| Polat | AIRS  AIRS with Fuzzy weighted pre-processing | 81.00  85.00 |
| Keles | ESTDD | 95.33 |
| Temurtas | MLNN with LM and  PNN  LVQ  MLNN with LM  PNN | 94.43  92.96  89.79  93.19  94.81 |
| Esin Dogantekina | GDA-WSVM | 91.86 |
| *Our method:* | SVM  KNN  Decision tree | 96.30  95.34  91.16 |

1. **Conclusion and Future Work**

Our paper gives a comparative analysis on diagnosis of thyroid disease by using various classification algorithms i.e. SVM, FKNN, and decision tree. After comparing the results that are shown in previous table, we found that SVM classifier technique provides better accuracies as compared to last works.

As a result following results can be summarised after working on these algorithms:

* After getting the accuracy of 96.30, it can be inferred that SVM could be successfully used for the diagnosis of thyroid disease.
* The SVM classifier used in this study shows better performance than the accuracies compared by Ozyilmaz and Yildirim, Polat, Keles and many other studies.

We can make a note that the proposed study performs efficiently on the dataset of 215 samples with an accuracy of 96.30. However, if we merge any other classification technique such as fuzzy classification or neural network on the output that we got from SVM algorithm, then the system might provide even better accuracy rate as compared to what we got with the current study. So for the future perspective we may follow the given paradigm to analyse the result for better accuracy for the diagnosing the thyroid disease.

1. **Acknowledgments**

I would like to express my special thanks of gratitude to my project guide Gopinath M.P. who gave us an opportunity to work on this project (Classification Of Algorithms For The Diagnosis Of Thyroid Disease) and also helped us in doing a lot of research in field of thyroid diagnosis which was totally new for us.

1. **References**
2. Halife Kodaz, Seral , Ahmet Arslan , Salih Gune: Medical application of information gain based artificial immune recognition system (AIRS): Diagnosis of thyroid disease.
3. G. Zhang, L.V. Berardi, An investigation of neural networks in thyroid function diagnosis, Health Care Manage. Sci. (1998)
4. Ali Keles, Ayturk Keles: ESTDD: Expert system for thyroid diseases diagnosis
5. Feyzullah Temurtas: A comparative study on thyroid disease diagnosis using neural networks.
6. Ozyılmaz, L., Yıldırım, T. (2002). Diagnosis of thyroid disease using artificial neural network

methods. In Proceedings of ICONIP’02 9th international conference on neural information processing (pp. 2033–2036). Singapore: Orchid Country Club.

1. ftp://ftp.ics.uci.edu/pub/machine-learning-databases, last accessed September 25, 2016.
2. Serpen, G., Jiang, H., and Allred, L., Performance analysis of probabilistic potential function neural network classifier. In Proceedings of artificial neural networks in engineering conference, St. Louis, MO, (Vol. 7, pp. 471–476), 1997.
3. Pasi, L., Similarity classifier applied to medical data sets, 2004, 10 sivua, Fuzziness in Finland’04. In International conference on soft computing, Helsinki, Finland & Gulf of Finland & Tallinn, Estonia, 2004
4. Polat, K. Sahan, & Gunes, S. (2007). A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted preprocessing for thyroid disease diagnosis. Expert Systems with Applications, 32, 1141–1147.
5. Keles, A., and Keles, A., ESTDD: Expert system for thyroid diseases diagnosis. Expert Syst. Appl. 34(1):242–246, 2008.
6. Temurtas, F., A comparative study on thyroid disease diagnosis using neural networks. Expert Syst. Appl. 36(1):944–949, 2009.
7. 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. 38(1):146–150, 2011.
8. Chen, H. L., Yang, B., et al., A three-stage expert system based on support vector machines for thyroid disease diagnosis.
9. Stegmayer, Georgina, Matias Gerard, and Diego H. Milone. "Data mining over biological datasets: An integrated approach based on computational intelligence."
10. Yeh, Wei-Chang. "Novel swarm optimization for mining classification rules on thyroid gland data." Information Sciences 197 (2012)
11. Chen, H. L., Yang, B., Wang, G., Liu, J., Chen, Y. D., & Liu, D. Y. (2012). A three-stage expert system based on support vector machines for thyroid disease diagnosis.Journal of medical systems, 36(3), 1953-1963.

[17] Design of an Enhanced Fuzzy k-nearest Neighbor Classifier Based Computer Aided Diagnostic System for Thyroid Disease Da-You Liu & Hui-Ling Chen & Bo Yang & Xin-En Lv & Li-Na Li & Jie Liu.

[18] K.Geetha, Capt. S. Santosh Baboo “Efficient thyroid disease classification using differential evolution with SVM”.

[19] Azar AT, Hassanien AE, Kim T (2012) Expert system based on neural fuzzy rules for thyroid diseases diagnosis. Springer, Berlin.