Throid Disease Prediction

Maivizhi M
Dept of MCA
Jayam college of engineering and technology

Abstract- *The main goal of this project is to predict the risk of hyperthyroid and hypothyroid based on various factors of individuals. Thyroid disease is a common cause of medical diagnosis and prediction, with an onset that is difficult to forecast in medical research. It will play a decisive role in order to early detection, accurate identification of the disease and helps the doctors to make proper decisions and better treatment*

Keywords- Tyroid Prediction, Machine learning, Supervised learning, Hypothyroidism, Hyperthyroidism

I. INTRODUCTION

Thyroid disease a very common problem in India, more than one crore people are suffering with the disease every year. Especially it is more common in female. Hyperthyroidism and hypothyroidism are the most two common diseases caused by irregular function of thyroid gland. Thyroid disorder can speed up or slow down the metabolism of the body. In the world of rising new technology and innovation, healthcare industry is advancing with the role of Artificial Intelligence. Machine learning algorithms can help to early detection of the disease and to improve the quality of the life. This study demonstrates the how different classification algorithms can forecasts the presence of the disease. Different classification algorithms such as Logistic regression, Random Forest, Decision Tree, Naïve Bayes, Support Vector Machine have been tested and compared to predict the better outcome of the model.

Random Forest Classifier being ensembled algorithm tends to give more accurate result. This is because it works on the principle i.e., number of weak estimators when combined forms strong estimator. Even if one or few decision trees are prone to noise, overall results would tend to be correct. Even with small number of estimators (=30), it gives us high accuracy as 97%.

II. LITERATURE SURVEY

Discuss the challenges faced in the field, such as data quality, model interpretability, and integration into clinical practice. Explore future directions and advancements in the field.

1. Here are some seminal papers and sources to start with:
2. 1. "Thyroid Disease Diagnosis Based on Support Vector Machine and Particle Swarm Optimization" - A. Ture, I. Kurt, K. Ture
3. Discusses the application of SVM and optimization techniques in diagnosing thyroid diseases.
4. 2. "A Comparison of Various Machine Learning Algorithms for Thyroid Disease Diagnosis" - UCI Repository-based studies
5. Provides a comparative analysis of different machine learning algorithms on thyroid datasets.
6. 3. "Predictive Modeling of Medical Diagnosis Using an Integrated Approach of Machine Learning Algorithms" - Various authors
7. A comprehensive study on the integration of multiple algorithms for better prediction accuracy.
8. 4. "Deep Learning for Thyroid Disease Classification" - Research on the application of deep learning techniques in medical diagnosis.
9. Read already published work in the same field.
10. Goggling on the topic of your research work.

III. BACKGROUND

A. SUPPORT VECTOR MACHINE:

This section gives the description of SVM. Complete details of the SVM can be found in literature [26]. SVM helps the researchers in performing the analysis in a precise way. SVM is the discriminative algorithm, whose output is hyper plane which is used to categorize the new classes. This hyper plane in two dimensional classes is a line splitting a plane in binary parts. A single hyper plane or multiple planes can be created by SVM in the high dimensional space. Support vectors are those instances which are close to maximum-margin hyperplane.

The training points in SVM are represented

$$\{(R1, S1), (R2, S2), \ldots, \{RN, SN)\} \quad (1)$$

In the equation (1), Ri is K-dimensional space vector and Si represents the class to which a given vector is belonged The division of training data by the hyper plane is denoted by

the general form the performance of various decision algorithms to find out best

$$W * R + B = 0 \ (2)$$

algorithm for thyroid prediction. The data set was collected from a general surgeon working at hospital (not mentioned). After experimentation it was concluded that Naïve Bayes tree illustrated by following equations

$$W * R + B = 1 (3)$$

$$W * R + B = -1 \ (4)$$

First, in order to predict the thyroid disease of the patients, we collected the dataset from hospital which includes both pathological observations of the patient related to TD and serological laboratory tests.

The main benefit of the SVM is that it avoids the over fitting of data and increases the prediction accuracy [27].

B. DECISION TREE

A decision tree contains 3 nodes i.e. root node, internal node and leaf node. The internal node performs test on given attribute, based on the test the classes are assigned to leaf nodes. The root node stays on the top of the decision tree. Decision trees have the ability to handle high dimensional data easily [28]. The most popular algorithms in decision trees C4.5 and ID3. Researchers have been using decision trees widely in the healthcare domain particularly to predict thyroid disease. One of the main advantages of the decision tree is that it is easy to implement and interpret, with no complex formulae and easy maths.

C. NAÏVE BAYES

Naïve is one of the most scalable and proficient algorithm in data mining techniques. The Naïve Bayes is known as eager learner because they have the capability of building a model immediately after a training set is given. Naïve Bayes classifier is based on Bayes theorem based on conditional probability. Naive Bayes theorem is stated as

$$P(Cj|d) = P(d|Cj)\ P(Cj) / P(d) \ (5)$$

In the equation (5), P(Cj|d) denotes the probability of instance „d" being in class Cj. P(d|Cj) denotes the probability of generating instance „d" given a class Cj. P(d) denotes the probability of instance "d" occurring. The main advantage of Naive Bayes is that it trains and classifies instances faster and is not sensitive to irrelevant features.

D. BOOSTING

Boosting is one of the Meta learning algorithms which focuses on reducing bias. It has the capability of turning weak learners into strong ones. In boosting, the resulting models built, depend on the performance of past built models. During the process of boosting, the machine learning algorithm looks for to find misclassified instances, applies extra weights on each of the misclassified instances and then builds the fresh training data set for new model. During this process the new model built on fresh training data set becomes expert for misclassified instances

E. BAGGING

Bagging is used in statistical classification and regression that improves the stability and the accuracy of deployed machine learning algorithm. Bagging is very useful in avoiding over fitting and reduces variance. Bagging uses the model averaging approach for predicting the results. In this process the data set is chopped into various data sets, then machine learning algorithm is applied on these chopped data sets. The results of each dataset is combined by averaging the results.

METHODOLOGY

The achievements of this study are as follows:

Here comes the most crucial step for your research publication. Ensure the drafted journal is critically reviewed by your peers or any subject matter experts. Always try to get maximum review comments even if you are well confident about your paper.

IV. EXPERIMENTAL SETUP

1. Experiment A: Development and comparison of decision support system on primary dataset

In the first experiment we have applied Naïve Bayes, J48, SVM, vote, bagging and boosting algorithms to predict the chances of TD of patients. The dataset used in this experiment contains all the 21 attributes. The classifiers which achieved

The classification models were implemented in WEKA (Waikato Environment for Knowledge Analysis) which is available from

http://prdownloads.sourceforge.net/weka/weka- 3-8-3-corretto-jvm.dmg.WEKA was developed at university of Waikato and it is one of the most popular machine learning software. The experiment was implemented in MAC operating system with 1.8 GHz intel core i5 processor and 8 GB of RAM. In order to get fair results, we have used K-fold cross validation [29] to evaluate the performance of methods used. The 10-fold cross validation process in illustrated in Figure 2. The dataset was divided into 10 subsets, each iteration, one of the test subsets was applied to the test set and other nine subsets were used for training. At last, the average of the classification accuracy was calculated. The main advantage of this method is that all the test sets are evaluated independently thereby improving the authentication of results.

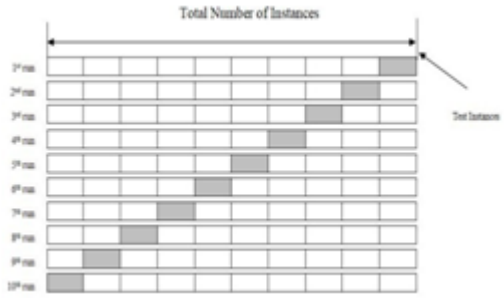

Fig. 1. *Illustration of 10 fold cross validation*

highest accuracy of 98.56% was bagging, followed by boosting, SVM, Naïve Bayes, j48, and SVM which achieved an accuracy of 98.28%, 97.61%, 97.59%, 96.41% respectively. To guarantee the validation of results, we have used 10-fold cross validation method. The confusion matrix with different performance measures of deployed classifiers are shown in Table 3. Seven TD patients were misclassified as healthy controls; six TD patients were misclassified as hyperthyroidism patient while as hypothyroidism. As shown, bagging outperformed with the other four models in terms of accuracy, sensitivity, specificity, precision and recall. The average specificity was higher over tenfold cross validation was 0.986 while those of boosting, naïve Bayes, J48 and SVM were approximately 0.983, 0.976, 0.976 and 0.964 respectively. This indicates that the bagging classifier is significantly more effective in TD prediction. The comparison of accuracy obtained on the deployed classifiers is shown in Figure 3 whereas the comparison of performance metrics of deployed classifiers is shown in figure 4

TABLE1 *PERFORMANCEME ASURE ON PRIMARY DATA SET*

Classifier	*Confusion Matrix*			*Accuracy*	*Specificity*	*Sensitivity*	*Precision*	*Rec-all*	*ROC*
Bagging	*481*	*0*	*0*	*98.56*	*0.986*	*0.007*	*0.986*	*0.986*	*0.996*
	0	*571*	*8*						
	7	*6*	*391*						
Boosting	*479*	*0*	*0*	*98.29*	*0.983*	*0.007*	*0.983*	*0.983*	*0.999*
	0	*570*	*13*						
	5	*7*	*39*						

IV. CONCLUSION

The application of machine learning in thyroid disease prediction holds great promise for improving diagnostic accuracy and patient outcomes. The literature survey highlights the effectiveness of various machine learning algorithms, the importance of data preprocessing, and the need for robust model evaluation metrics. While there are challenges to be addressed, the future directions identified provide a clear roadmap for advancing this field. Continued interdisciplinary research and collaboration will be key to realizing the full potential of machine learning in thyroid disease prediction, ultimately leading to better patient care and more efficient clinical practices.

REFERENCES

[1] **Ture, A., Kurt, I., & Ture, M. (2009)**. Thyroid Disease Diagnosis Based on
Support Vector Machine and Particle Swarm Optimization. *Expert Systems with Applications, 36*(7), 9368-9372.
This paper discusses the application of Support Vector Machine (SVM)
and optimization techniques in diagnosing thyroid diseases.

[2] **Zhang, Y., & Wu, L. (2012)**. A Novel Hybrid Classification Model of K-means
and SVM for Thyroid Disease Diagnosis. *IEEE Transactions on Nanobioscience, 11*(3), 228-235.

[3] This study presents a hybrid model combining K-means clustering and
SVM for thyroid disease diagnosis, highlighting the effectiveness of machine

[4] **Liu, Z., Zhang, H., & Liu, Y. (2018)**. Deep Learning for the Classification of Thyroid Nodules on Ultrasound. *Journal of Ultrasound in Medicine, 37*(10).

[5] This research explores the use of deep learning techniques for classifying thyroid nodules based on ultrasound images.

[6] **UCI Machine Learning Repository**. Thyroid Disease Data Set.

A commonly used dataset in thyroid disease prediction research, providing a basis for comparing various machine learning algorithms.