Predicting Hypothyroid Disease using Ensemble Models through Machine Learning Approach

Mohammed Ali Shaik[1], Praveen Pappula[2] and T Sampath Kumar[3]

Author Affiliations
[1,2,3] *Department of Computer Science &Artificial Intelligence, SR University, Warangal, Telangana State, India.*

Author Emails
[1)] niharali@gmail.com
[2)] prawin1731@gmail.com
[3)] tsk0707@gmail.com

Abstract. In the present era most of the people are effected with hypothyroid disease. People who are of age 13 and above are more effected with this disease and day by day it is transforming into a dangerous disease. The prediction of disease at earlier stage is very crucial so superior treatment is contributed by doctors. In this paper, hypothyroid disease prediction is implemented through Ensemble Machine Learning Algorithms and through RapidMiner Tool. The algorithms like Random Forest, AdaBoost, and Gradient Boosting and Bagging are implemented. These models were compared and evaluated using evaluation metrics like Accuracy, Precision, and Recall. As multiple models are combined so ensemble method is more robust and accurate. It removes unrelated data in the medical dataset and produces accurate and precise data. It proves that the ensemble method is very efficient than using a single classifying method. Hence implementing the ensemble method, we can predict the patient's disease efficiently.

INTRODUCTION

The "Hypothyroidism" is a provision where the "thyroid gland" which cannot produce enough "thyroid hormones" which slowdowns metabolism. It is stage where the levels of Triodothyronine (T3) and Thyroxin (T4) reduces and the level of Thyroid Stimulating Hormone (TSH) increases. Hypothyroidism can cause severe obstacles if not found early. Diagnosing hypothyroid disease is not an easy task as it includes many steps. The ordinary established method is a proper health checkup and many blood tests.

In the medical field, machine learning plays a vital role in diagnosing hypothyroid disease as it has several classification models based on which the model is trained with a dataset of the hypothyroid patient and predicts disease. Based on the values obtained from predictions, the medical staff easily checks the patient's condition and disregards further laboratory tests. The prediction model decreases the price of treatment and saves time for thyroid patients. Therefore, it is very favorable to the health maintenance sector.

In this paper, the dataset used is downloaded from the UCI repository is used. The whole work is accomplished with RapidMiner Tool, which is an open-source software.

LITERATURE SURVEY

As per reference [1] has implemented Decision Tree and Naive Bayesian classifiers on the thyroid dataset which was downloaded from Kaggle. They initially trained the dataset and applied the Decision Tree algorithm which generates yes or no values as result. If this value is yes, then the Naive Bayesian classifier is implemented to calculate the stage of thyroid disease. If the value is no, then it represents the patient doesn't have thyroid disease. They have concluded that this process reduces the complex work and cost.

As per reference [2] has proposed the method which consists of 3 stages- Data preprocessing, Feature Dimension Reduction, and Classification & Reduction. Firstly, they divided the dataset into k-different subsets. Next, they applied sampling techniques on subsets and performed Principal Component Analysis (PCA) transformation on them, and constructed a rational matrix to produce new feature space. Finally, they reduced the feature space and implemented Random Forest. They have concluded that their proposed method achieves high accuracy.

As per reference [3] the process of feature selection methods like Principal Component Analysis (PCA), RFE, and UFS accompanied by ML algorithms like RF, LR, SVM, and DT. They have concluded that usage of RFE enhanced the accuracy of algorithms.

As per reference [4] has implemented four algorithms like Artificial Neural Network (ANN), Naïve Bayes, SVM, and DT on the dataset. They have evaluated the performance of the models using Accuracy and Mean Absolute Error. They concluded that among these algorithms, the SVM classifier has the highest accuracy and least Mean Absolute Error.

As per reference [5] has implemented a three-class classification model through Decision Tree, KNN, and Naïve Bayes algorithms. They have used 10-fold cross validation and calculated the performance using Evaluation Metrics. They have concluded that the Decision Tree gives the most accurate results.

As per reference [6] has implemented Naïve Bayes, KNN, SVM, and ID3 Algorithms on the thyroid. They have interpreted the correlation between the attributes T3, T4, TSH, and gender in the patients' dataset. They evaluated algorithms according to their accuracy, and speed.

As per reference [7] has proposed a methodology using two CART decision trees. Firstly, they implemented feature rejection and then assigned costs for training decision trees. Later an algorithm that is based on exponential weights was used for weighing each classifier. They have concluded that their method is effective because weights are updated continuously to enhance the accuracy of the model.

As per reference [8] who has proposed a Multi-Class Support Vector Machine (MCSVM) Classifier to detect the stage of thyroid disease. Evaluation metrics used were Accuracy, Precision, and Recall. He concluded that the MCSVM gives more accurate results in the detection of four different stages of the thyroid.

As per reference [9] have proposed Univariate Selection, TreeBased Feature Selection, and Recursive Feature Elimination as feature selection methods. They have implemented SVM, Random Forest, and Naïve Bayes classification techniques. They have concluded that SVM gives accurate results and can be used to classify symptoms of hypothyroid disease.

As per reference [10] has used two algorithms Multilayer Perceptron (MLP) and Naïve Bayes for classification. They have implemented on WEKA Tool and analyzed the performance of algorithms. They have concluded that Naïve Bayes gives the highest accuracy.

As per reference [11] has used SVM for classification. They have implemented Particle Swarm Optimization to optimize SVM parameters. In the testing phase, they used KNN imputation to approximate the missing values.

As per reference [12] has used four classification algorithms, Logistic Regression, Artificial Neural Network, KNN, and SVM to identify the level of the disease. According to the results, they have concluded that LR gives the highest accuracy but after parameter tuning and standardizing the data, SVM was found the best.

As per reference [13] has carried out the entire work using WEKA Tool. They have implemented Ranker Search as a Feature Optimization method and a machine learning algorithm-Naïve Bayes Classifier. They have analyzed the performance of the classifier using Root Mean Square Error, Accuracy, Precision, Recall, Systems build-up time comparative analysis, and F-Measure.

As per reference [14] have proposed a model named Multi Kernal Support Vector Machine (MKSVM). They used an optimal feature selection and evaluated the performance of a model using measures like Accuracy, Specificity, and Sensitivity. The accuracy of the model achieved was 97.49%.

As per reference [15] has implemented six methods like RBF, LVQ, MLP, BPA, AIRS, and Perceptron. They have used measures like accuracy, precision, recall, and f-measure for performance calculation. They have concluded that Multilayer Perceptron (MLP) has the highest accuracy among others.

PROPOSED METHODOLOGY

The proposed methodology involves four stages: Data Pre-processing, Feature Selection, Ensemble Machine Learning Models implementation, and Predictions. Ensemble models mean the combination of different models. The data flow of the proposed method is as follows Fig.1.

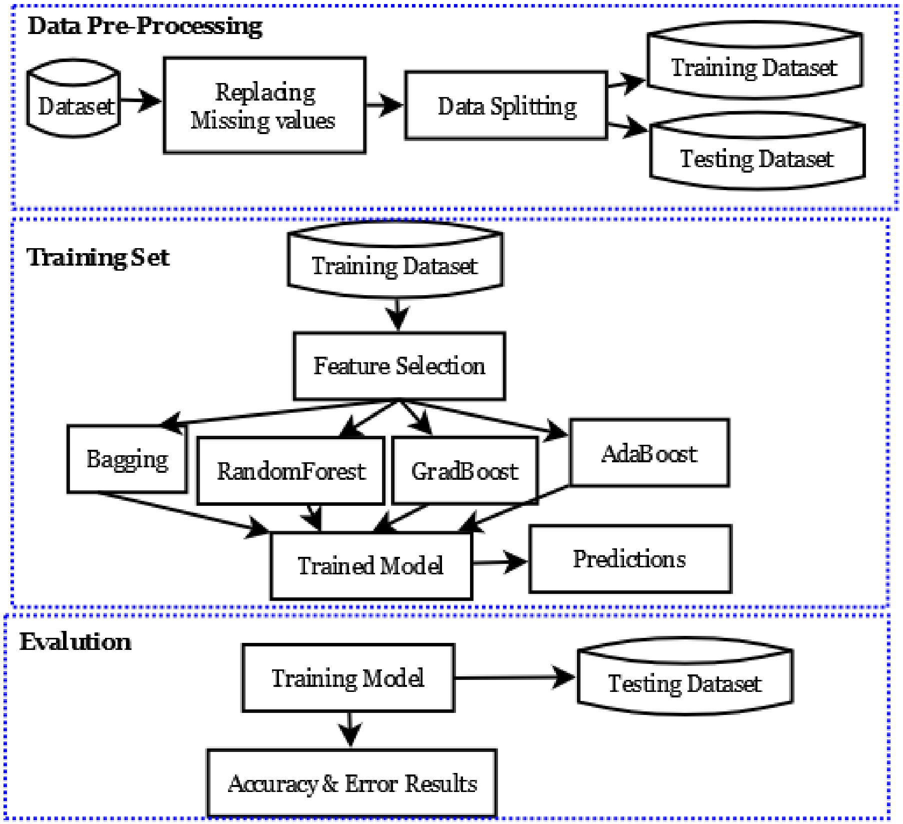

FIGURE 1. The framework of our proposed model

Dataset

The dataset is an assortment of related information. The dataset has been downloaded from the UCI repository. It has a record of patients' data which states whether they are suffering from hypothyroidism or not. It has 28 features, they are age, gender, symptoms of hypothyroid and the feature which addresses the label is “binary class”. The binary class takes only 2 values: if the value is 1 then hypothyroid is positive and if the value is 0 then it is hypothyroid is negative.

Data Preprocessing

Data preprocessing is implemented to make over the raw data comprising of “noisy and missing data” consists of functional and competent format. It formats the data and makes it applicable for machine learning models.

Feature Selection

It is the method of minimizing several input variables to improve the performance of the predictive model while developing it. It reduces the computational cost, removes unwanted features, and makes training faster. The Chi-

Square Test method is used for feature selection. This evaluates the attribute weights over each of the class attribute by utilizing the chi-squared statistics that weight that higher attribute which is treated as the most relevant attribute. It is used to calculate only nominal labels. It uses frequencies instead of means and variances. The value is calculated as:

$$X^2 = \sum \left[\frac{(O-}{E} \right. \tag{1}$$

Where in equation 1 X^2 is chi square statistics, then O is denoted as frequency and E denotes expected frequency.

Data Splitting

While splitting the data, two independent variables are defined X and Y. where the dataset is divided into dual forms “Training set and a Testing set” in the ratio of 70:30 respectively. These two sets are used for the further evaluation process.

Algorithms

There are different Machine Learning Algorithms to predict the medical diagnosis data. In this project, three ensemble algorithms “Random Forest, AdaBoost, and Gradient Boosting” were used to forecast the disease and are compared using evaluation metrics.

- **Random Forest**: algorithm is also called as Random Decision Forest Algorithm. This method is based on supervised learning and is implemented in both “Classification and Regression” problems. It is an ensemble method and includes numerous constructions of decision trees to perform various tasks. Each tree generates its respective results. These results are integrated simultaneously to urge more accurate predictions. This method gives better accuracy results compared to other existing algorithms.
- **Gradient Boosting**: it is a “supervised learning algorithm” and is used in performing both “Classification and Regression problems”. In this algorithm, the base model is created and predictions are made. The residuals formed from the model are transferred to another new model as the target. This process continues ‘n’ no. of times. As numerous models are involved in this algorithm it is the Ensembled method.
- **AdaBoost**: AdaBoost Technique is also termed Adaptive Boosting. Most commonly it is used with a Decision Tree of only one level (1 Split) which are labeled Decision stumps. In this algorithm, the base model is built and weights are assigned equally for each datapoint. After training the model, the higher weights are allotted to errors. Now, in the next model, errors with the highest weights are given more importance. This process continues till the least errors are received. This algorithm is also used as an ensemble method because it uses many models.

EVALUATION METRICS

Different types of measures are used to evaluate algorithms. They are “Accuracy, Precision, Recall, and Confusion Matrix”. These metrics are calculated using “True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN)”.

Accuracy

The accuracy score is the ratio of no. of correct predictions to the total number of interpretation

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+} \tag{2}$$

Precision

The “precision” is the ratio of accurately predicted with “positive observations” to the total “predicted positive observations”

$$Precision = \frac{T.}{(TP+} \quad (3)$$

Recall

The recall is the “ratio of correctly predicted positive observations to the sum of predicted positive and negative observations”

$$Recall = \frac{T.}{(TP+} \quad (4)$$

Predictions

It is the process of detecting the output or some patterns from the existing past data which is used to make future decisions and based on the results appropriate decisions can be taken.

EXPERIMENTAL SETUP

RapidMiner Tool is used for implementing this model. This tool has different types of operators for each task. These operators help us to build a model. Firstly, the dataset should be imported into the tool. The next dataset is retrieved using the Retrieve operator and data is pre-processed using Replacing Missing Values operator. Then Chi-square test feature selection is applied using the Chi-Squared Statistic operator. After that, the dataset is divided into “training and testing sets” in the ratio of 75:25. The algorithms are implemented on the “training set”. The accomplishment of the model is estimated by evaluating the metrics. Finally, the predictions of the algorithm which has the highest accuracy is only considered.

RapidMiner Tool

It is an impressive “data mining tool” that implements whole from “data mining” to model and further deploy the model procedure by offering the incorporated environment for data generation then performing “predictive analytics by imposing the machine learning capabilities and deep learning”. It is used for “education, research, training, and application development”. It even includes “visualization, model validation, and optimization” as it performs “services Backward Engineering”.

RESULTS

Based on the results obtained by implementing all the four algorithms, the accuracy of Random Forest is higher than other algorithms. So that the predictions made by Random Forest are considered.

TABLE 1. Random Forest Implementation results

Random Forest	True:1	True: 0	Class Precision
Prediction:1	2460	0	100%
Prediction:0	9	223	97%
Class Recall	99.8%	100%	

Table 1 represents the results obtained by Random Forest and the accuracy obtained is 99.75 percentage

TABLE 2. Gradient Boosting Implementation results

Gradient Boosting	True:1	True: 0	Class Precision
Prediction: 1	1045	5	99.7%
Prediction: 0	9	85	90.2%
Class Recall	99.3%	95.3%	

Table 2 represents the results obtained by Gradient Boosting and the accuracy obtained is 98.97 percentage

TABLE 3. Ada Boost Implementation results

Ada Boost	True:1	True: 0	Class Precision
Prediction: 1	1025	16	98.6%
Prediction: 0	26	75	76.2%
Class Recall	97.8%	84.3%	

Table 3 represents the results obtained by Ada Boost and the accuracy obtained is 96.67 percentage

TABLE 4. Bagging Implementation results

Bagging	True:1	True: 0	Class Precision
Prediction: 1	1048	3	98.6%
Prediction: 0	6	88	97.8%
Class Recall	97.8%	84.3%	

Table 4 represents the results obtained by Bagging and the accuracy obtained is 99.53 percentage and Table 5 represents the comparative study of all the algorithms being implemented.

TABLE 5. Comparative study of algorithms implemented

Algorithm	Accuracy (%)
Random Forest	99.75
Gradient Boosting	98.97
Ada Boost	96.67
Bagging	99.53

CONCLUSION AND FUTURE SCOPE

Diagnosing the hypothyroid disease manually is a complicated task, as the “machine learning algorithms” were it is used to perform predicting the disease efficiently. The “ensemble machine learning” methods like “Random Forest, Gradient Boosting, AdaBoost, and Bagging” are used in this project to get accurate results. The performance of the model is high when the feature selection is added to the “trained model”.

The accuracy of Random Forest is higher than Gradient Boosting AdaBoost, and Bagging. Hence the Ensemble ML Models are very efficient and accurate in the prediction of hypothyroid disease. Moreover, extensive hyper parameter tuning of ML and upgraded feature selection would be performed for superior performance. This method is very favorable to the health maintenance sector.

REFERENCES

1. M. A. Shaik, S. k. Koppula, M. Rafiuddin and B. S. Preethi, "COVID-19 Detector Using Deep Learning," 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), 2022, pp. 443-449, doi: 10.1109/ICAAIC53929.2022.9792694.

2. T.Sampath Kumar, B.Manjula, "Security Issue Analysis on Cloud Computing Based System",International Journal of Future Generation Communication and Networking Vol. 12, No. 5, (2019), pp. 143 - 150
3. Mohammed Ali Shaik and Dhanraj Verma, "Prediction of Heart Disease using Swarm Intelligence based Machine Learning Algorithms", International Conference on Research in Sciences, Engineering & Technology, AIP Conf. Proc. 2418, 020025-1–020025-9; https://doi.org/10.1063/5.0081719, Published by AIP Publishing. 978-0-7354-4368-6, pp. 020025-1 to 020025-9
4. YerrollaChanti, Bandi Bhaskar, NagendarYamsani, “Li-Fi Technology Utilized In Leveraged To Power In Aviation System Entertainment Through Wireless Communication”, J. Mech. Cont.& Math. Sci., Vol.-15, No.-6, June (2020) pp 405-412.
5. Mohammed Ali Shaik and Dhanraj Verma, "Predicting Present Day Mobile Phone Sales using Time Series based Hybrid Prediction Model", International Conference on Research in Sciences, Engineering & Technology, AIP Conf. Proc. 2418, 020073-1–020073-9; https://doi.org/10.1063/5.0081722, Published by AIP Publishing. 978-0-7354-4368-6, pp. 020073-1 to 020073-9
6. Mohammed Ali Shaik, MD.Riyaz Ahmed, M. Sai Ram and G. Ranadheer Reddy, "Imposing Security in the Video Surveillance", International Conference on Research in Sciences, Engineering & Technology, AIP Conf. Proc. 2418, 020012-1–020012-8; https://doi.org/10.1063/5.0081720, Published by AIP Publishing. 978-0-7354-4368-6, pp. 020012-1 to 020012-8.
7. Mohammed Ali Shaik, Geetha Manoharan, B Prashanth, NuneAkhil, Anumandla Akash and Thudi Raja Shekhar Reddy, "Prediction of Crop Yield using Machine Learning", International Conference on Research in Sciences, Engineering & Technology, AIP Conf. Proc. 2418, 020072-1–020072-8; https://doi.org/10.1063/5.0081726, Published by AIP Publishing. 978-0-7354-4368-6, pp. 020072-1 to 020072-8
8. Mohammed Ali Shaik and Dhanraj Verma, (2020), Enhanced ANN training model to smooth and time series forecast, 2020 IOP Conf. Ser.: Mater. Sci. Eng. 981 022038, doi.org/10.1088/1757-899X/981/2/022038
9. T. Sampath Kumar, B. Manjula, Mohammed Ali Shaik, Dr. P. Praveen, (2019), A Comprehensive Study on Single Sign on Technique International Journal of Advanced Science and Technology (IJAST) Vol-127 pp 156-162.
10. Mohammed Ali Shaik, Dhanraj Verma, P Praveen, K Ranganath and Bonthala Prabhanjan Yadav, (2020), RNN based prediction of spatiotemporal data mining, 2020 IOP Conf. Ser.: Mater. Sci. Eng. 981 022027, doi.org/10.1088/1757-899X/981/2/022027
11. T. Sampath Kumar,B. Manjula, D. Srinivas, (2017), A New Technique to Secure Data Over Cloud Jour of Adv Research in Dynamical & Control Systems vol 11 pp 145-149.
12. Mohammed Ali Shaik and Dhanraj Verma, (2020), Deep learning time series to forecast COVID-19 active cases in INDIA: A comparative study, 2020 IOP Conf. Ser.:Mater.Sci.Eng. 981 022041, doi.org/10.1088/1757-899X/981/2/022041
13. P Praveen, M Ranjith Kumar, Mohammed Ali Shaik, R Ravi kumar and R Kiran, (2020), The comparative study on agglomerative hierarchical clustering using numerical data, 2020 IOP Conf. Ser.: Mater. Sci. Eng. 981 022071, doi.org/10.1088/1757-899X/981/2/022071
14. P.Praveen, B.Rama, (2019), An Efficient Smart Search Using R Tree on Spatial Data Journal of Advanced Research in Dynamical and Control Systems vol 4 pp1943-1949.
15. Mohammed Ali Shaik, “Time Series Forecasting using Vector quantization”, International Journal of Advanced Science and Technology (IJAST), ISSN:2005-4238,Volume-29,Issue-4 (2020), Pp.169-175.
16. Mohammed Ali Shaik, T. Sampath Kumar, P. Praveen, R. Vijayaprakash, “Research on Multi-Agent Experiment in Clustering”, International Journal of Recent Technology and Engineering (IJRTE), ISSN:2277-3878,Volume-8,Issue-1S4, June2019. Pp. 1126-1129.
17. Mohammed Ali Shaik, “A Survey on Text Classification methods through Machine Learning Methods”, International Journal of Control and Automation (IJCA), ISSN:2005-4297,Volume-12,Issue-6 (2019), Pp.390-396.

18. R. Ravi Kumar, M. Babu Reddy and P. Praveen, (2017), A review of feature subset selection on unsupervised learning Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), pp163-167