Human-Intelligent Systems Integration
https://doi.org/10.1007/s42454-020-00006-y

RESEARCH ARTICLE

Prediction of thyroid disease using decision tree ensemble method

Dhyan Chandra Yadav[1] · Saurabh Pal[1]

Received: 18 June 2019 / Accepted: 12 March 2020
© Springer Nature Switzerland AG 2020

Abstract
Thyroid disease is spreading very rapidly among women after the age of 30 years. Therefore, it is necessary to examine the thyroid dataset for predicting the disease at early stage so that precautions can be taken to protect the dangerous condition of thyroid cancer. A decision tree is used to extract hidden patterns from the stored datasets. The objective of this research paper is to examine the thyroid disease dataset using decision tree, random forest, and classification and regression tree (CART), and after obtaining the results of these classifiers, we enhanced the results using the bagging ensemble technique. The proposed experiment was done on 3710 instances and 29 features of thyroid patients. The overall prediction depends on target variable whch is divided in sick and negative class. The accuracy of the prediction was calculated on the basis of different num-fold and seed values. Different classification algorithms are analyzed using thyroid dataset. The results obtained by individual classification algorithms like decision tree, random forest tree, and extra tree give an accuracy of 98%, 99%, and 93%, respectively. Then, we developed a bagging ensemble method combining the three basic tree classifiers and apply again on the same dataset, which gives a better accuracy of 100% in the case of seed value 35 and num-fold value 10. This proposed ensemble method can be used for better prediction of thyroid disease.

Keywords Decision tree · Random forest · Extra tree and boosted tree ensemble model · Triiodothyronine · Thyroxin and thyroid stimulating hormone

1 Introduction

Hormones play a vital role in blood flow for maintaining metabolism in human, and high hormones and low hormones both are equally dangerous. Thyroid hormones are produced by thyroid gland to maintain blood stream for the regulation of metabolism; three types of hormones produced by the thyroid gland are triiodothyronine (T3), thyroxin(T4), and thyroid-stimulating hormone (TSH). If the thyroid gland produces more hormones, then, it will be hyperthyroidism, and if the thyroid gland produces less hormones, then, it will be hypothyroidism. Thyroid disease has various symptoms, like fatigue, weakness, intolerance to cold, muscle aches and crams, constipation, weight gain, or difficulty of losing weight, in initial stage; therefore, it is necessary to recognize thyroid disease in the initial stage (Ozyilmaz and Yildirim 2002).

Tahani et.al used adaptive clustering ensemble model and combine multiple clustering models for prediction of thyroid disease. Adaptive clustering method computed and transformed initial clusters into binary representation to predict final clusters using K-means algorithm (Alqurashi and Wang 2019).

Ahmad et.al discussed feature selection techniques and achieved different testing phase of clustering one, two, three, and four for each class and 12 fuzzy rules to calculate the maximum absolute difference, linguistic hedge, and total serum thyroxin. They achieved classification accuracy of 98.60% (Azar et al. 2012).

Xiyu et.al analyzed traditional tissue P systems to generate new class of tissue system. They used thyroid disease analysis, tissue P system, membranes structure, and clustering algorithm for the prediction of thyroid disease using classification algorithms (Liu and Xue 2012).

✉ Dhyan Chandra Yadav
dc9532105114@gmail.com

Saurabh Pal
drsaurabhpal@yahoo.co.in

[1] Dept. of Computer Applications, VBS Purvanchal University, Jaunpur, India

Published online: 06 April 2020

Ahmad et.al analyzed thyroid disease using compression hard and fuzzy clustering methods and found optimal number of clusters. They used K-means, K-model clustering fuzzy C-means for prediction of thyoid disease. They improved the actual number of clusters present in thyroid dataset and find that clustering performance is much as compared with other (Azar et al. 2013).

Vikas et.al discussed different machine learning algorithms for the analysis of chronic kidney disease. They used different data mining algorithms as a decision tree and regression tree and CART. They found the prediction accuracy of 93%. They also suggested boosting ensemble technique for prediction (Chaurasia et al. 2018a).

Awasthi and Anil Antony discussed about the classification and diagnosis of thyroid disease using KNN, support vector machine (SVM), and machine learning algorithms. They used K-nearest neighbor algorithm in thyroid diagnosis for approximating the missing values in the user input (Aswathi and Antony 2018).

Vikas et.al discussed about breast cancer in women using Naïve Bayes and RBF network machine learning algorithms. They predicted that Naïve Bayes gives the highest accuracy of 97.36% (Chaurasia et al. 2018b).

Yadav and Pal discussed thyroid disease prediction using decision tree, overfitting, and neural network machine learning algorithms. They used AdaBoost, bagging, boosting, and stacking ensemble techniques to enhance the predicted values. They got bagging and boosting ensemble techniques combined as the best and with an accuracy of 98.79% (Yadav and Pal 2019).

2 Methodology

Figure 1 illustrates the methodology used in this research paper. First, we choose the thyroid disease dataset, and then, the dataset is analyzed using three data mining techniques, decision tree, random forest, and extra tree. A bagging ensemble technique is then used to combine the result obtained by the three different classifiers to enhance the prediction values. Finally, we get the best predicted values for analyzing the thyroid disease.

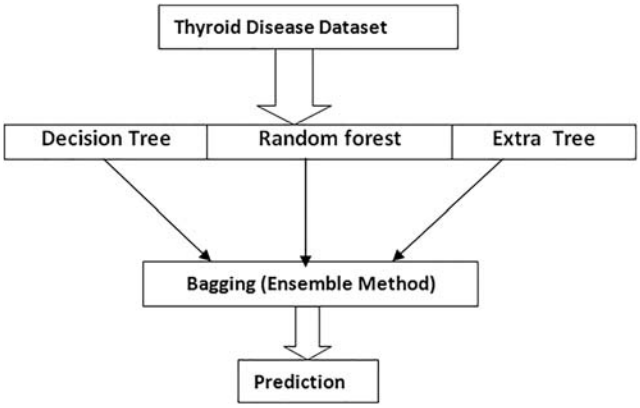


Fig. 1 Proposed methodology

2.1 Data description

Thyroid disease dataset is taken from the UCI machine learning repository for analysis. The original dataset consists of 3710 instances and 30 features. In the preprocessing step, we leave one feature (sex) and fill the missing values using the moving average method. Table 1 shows the list of 29 features, there types, and range of each feature. The decision variable class is divided into two types: positive and negative. The values of features beyond the medically prescribed limit

Table 1 Description for thyroid disease dataset (http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease 2013)

Class	Features	Description	Domain
Positive	Age	Real	[0.01, 0.97]
Negative	On_thyroxine	Integer	[0, 1]
	Query_on_thyroxine	Integer	[0, 1]
	On_antithyroid_ medication	Integer	[0, 1]
	Sick	Integer	[0, 1]
	Pregnant	Integer	[0, 1]
	Thyroid_surgery	Integer	[0, 1]
	I131_treatment	Integer	[0, 1]
	Query_hypothyroid	Integer	[0, 1]
	Query_hyperthyroid	Integer	[0, 1]
	Lithium	Integer	[0, 1]
	Goiter	Integer	[0, 1]
	Tumor	Integer	[0, 1]
	Hypopituitary	Integer	[0, 1]
	Psych	Integer	[0, 1]
	TSH_ measured	Integer	[0,1]
	TSH	Real	[0.0, 0.53]
	T3_measured	Integer	[0,1]
	T3	Real	[0.0005, 0.18]
	TT4_measured	Integer	[0,1]
	TT4	Real	[0.0020, 0.6]
	T4U_measured	Integer	[0,1]
	T4U	Real	[0.017, 0.233]
	FTI_measured	Integer	[0,1]
	FTI	Real	[0.0020, 0.642]
	TBG_measured	Integer	[0,1]
	TBG	Real	[1.1–2.1]
	Referral Source	Integer	[0,1,2,3,4]

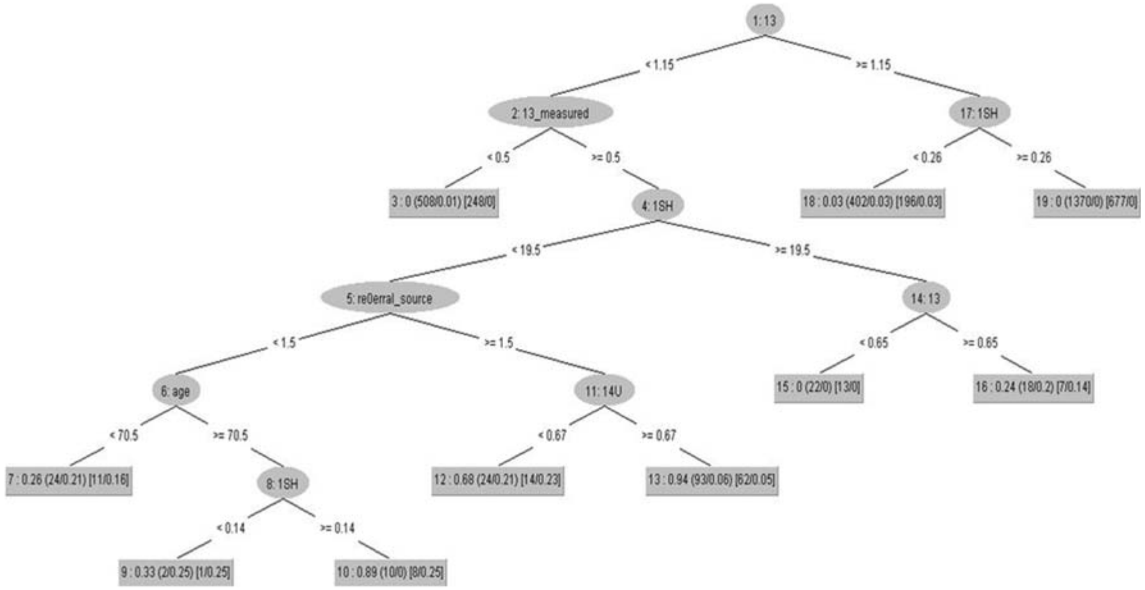


Fig. 2 Decision tree representation for thyroid dataset

are sick patients treated as positive and whose limit is within the range are negative. The alphabetic values are changed into numerical values, 0 for false and 1 for true.

2.2 Classifiers description

Three classifiers, decision tree, random forest, and extra tree, are used for calculating the prediction of thyroid disease. Brief descriptions of these classifiers are as follows:

Decision tree The decision tree easily divides instances and features. We easily take decision in prediction by decision tree and find the estimation of outcomes and take decision for future planning in any medical diagnosis or in other areas.

In the decision tree, attributes are divided into subnode as decision node and, by the help of machine learning tools easily, represent thyroid dataset in tree form as shown below in Fig. 2.

Random forest tree A random forest tree is a forest of trees in which many trees support to take decision in prediction. It provides the best split of all attributes of medical data or other areas. This tree generates ideas for constructing many learners in machine learning, and by the help of

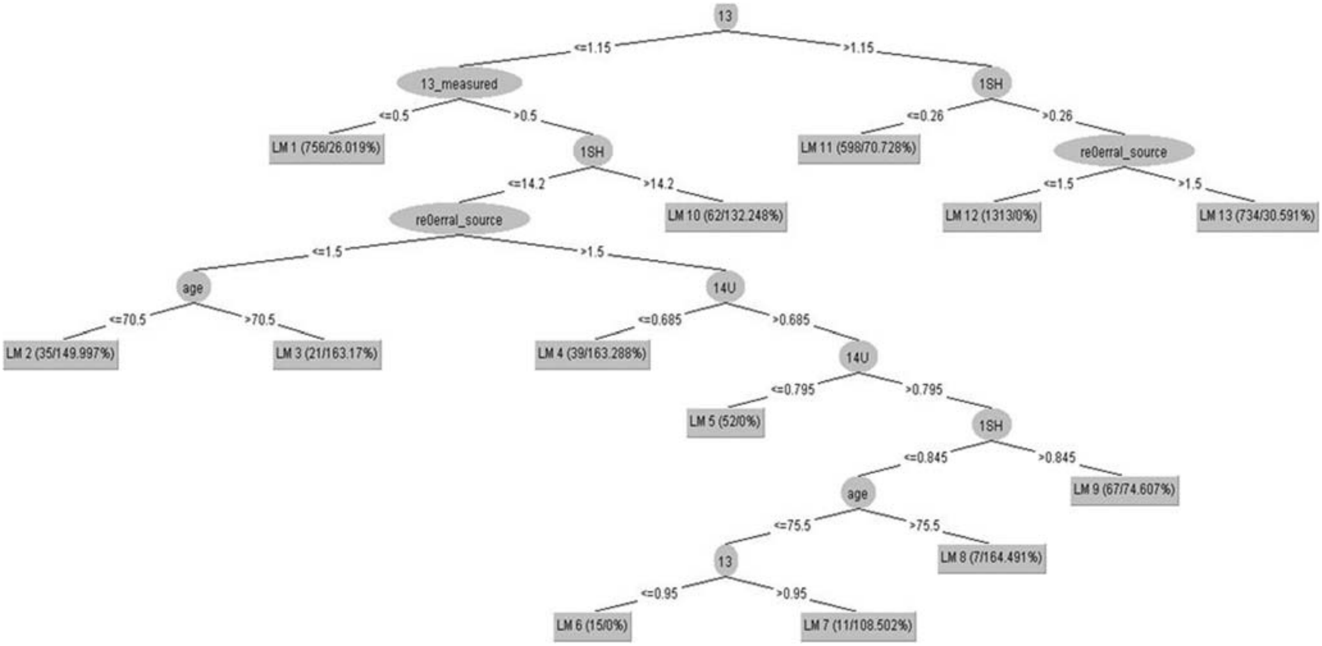
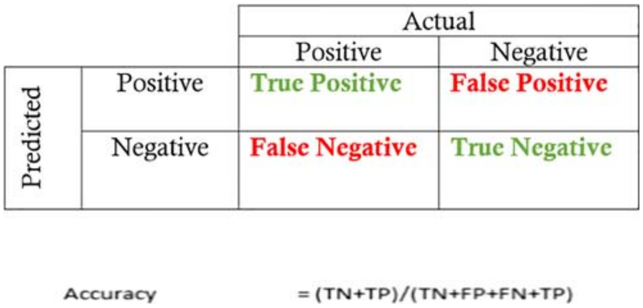


Fig. 3 Random tree representation for thyroid dataset

machine learning tool, we easily transform thyroid dataset as shown below in Fig. 3.

Extra tree The extra tree is a machine learning tree, and it provides help in making decision by split node attributes in random format. In this research paper, we observe how these classifiers are different and similar to each other and use thyroid dataset for best prediction. It is also easily useable for different areas of dataset [http://www.irdindia.in/journal_ijaece/pdf/vol3_iss4/2.pdf, (Landwehr et al. 2005; Breiman 2001; Sharaff and Gupta 2019)].

Ensemble method In this research paper, we use an ensemble method to enhance the results obtained from decision tree, random forest, and extra tree. We use bagging ensemble method to measure their performance in single unit for best prediction. The bagging ensemble method performs prediction in a parallel way on training, and on testing data also, this classifier does not avoid data variance. In this paper, bagging ensemble method will predict on thyroid dataset and measure classification accuracy with confusion matrix.



		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Accuracy = (TN+TP)/(TN+FP+FN+TP)

3 Results and discussion

Thyroid dataset which consists of 3710 instances and 29 features is visualized with the help of box and whisker plot as shown in Fig. 4. Each feature is described by its values, and the 50th and 75th percentile shown by the box and middle line in the box shows the mean of the values of that features. All 28 features are represented with the help of box and whisker to easily understand the mean and percentile of that attributes.

In all the observation, we take different seed values, but in this condition, num-fold values remain constant, and in the second phase, we take num-fold value as dynamic, and seed values remain constant. Seed and num-fold both are two accuracy improvement supporting ideas. Seed generates random number during process support to improve prediction, accuracy and num-fold performance to repeat the calculation in entire dataset and pick the classifier for better performance with different number of seed. In this research paper we test and measure thyroid dataset and their features' performance. The seed and num-fold values and accuracy, confusion matrix for different classifiers, and bagging ensemble method are shown in Table 2.

Confusion matrix is a matrix of true positive, true negative, false positive, and false negative in a tabular form. Confusion matrix is basically used for knowing the true value for test data and used to visualize their performance.

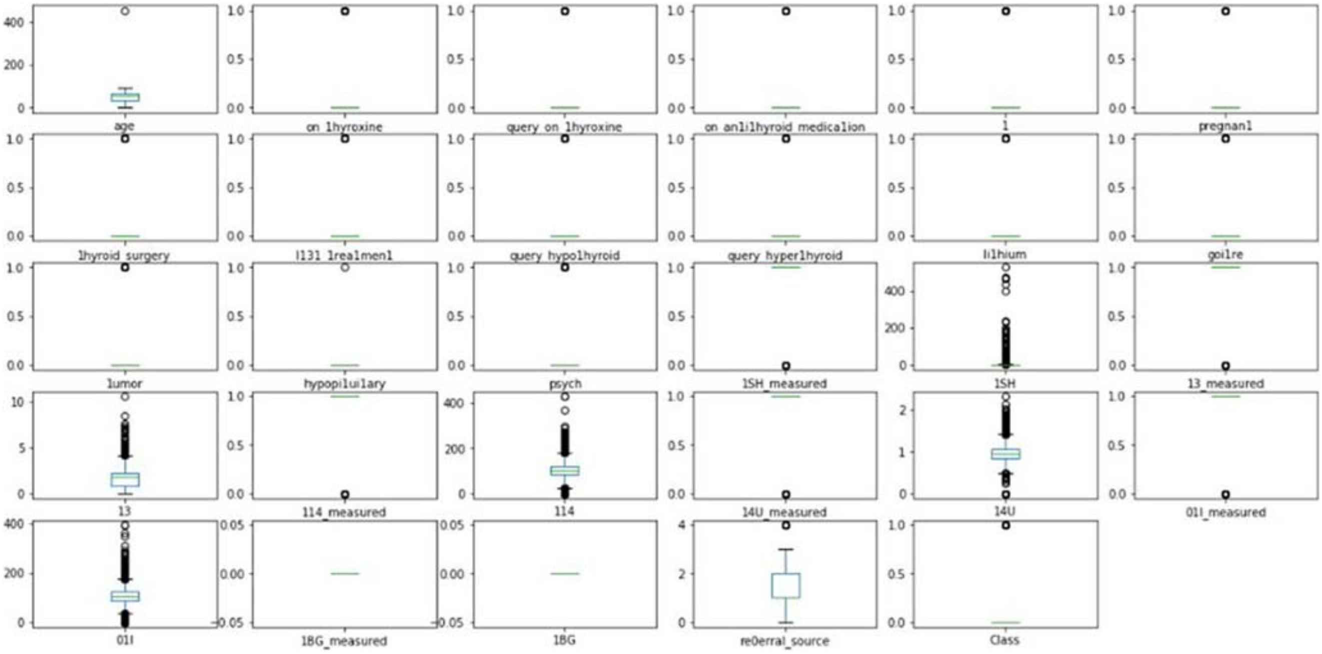


Fig. 4 Box and whisker plot of thyroid dataset features

Table 2 Computational table for thyroid dataset using different classifier

Iterations	Num folds	Seed	Accuracy with confusion matrix			
			DTC	RFC	ETC	BAGG
1	10	1	Acc:0.98 Conf: [[697 3] [6 36]]	Acc:0.97 Conf: [[698 2] [13 29]]	Acc:0.94 Conf: [[684 16] [25 17]]	Acc:0.99 Conf: [[700 0] [6 36]]
2	10	5	Acc:0.98 Conf: [[687 7] [1 47]]	Acc:0.98 Conf: [[692 2] [8 40]]	Acc:0.92 Conf: [[661 33] [24 24]]	Acc:0.98 Conf: [[690 4] [4 44]]
3	10	10	Acc:0.98 Conf: [[688 1] [11 42]]	Acc:0.97 Conf: [[687 2] [16 37]]	Acc:0.94 Conf: [[673 16] [23 30]]	Acc:0.98 Conf: [[689 0] [10 43]]
4	10	15	Acc:0.98 Conf: [[675 10] [4 53]]	Acc:0.97 Conf: [[679 6] [15 42]]	Acc:0.94 Conf: [[668 17] [26 31]]	Acc:0.98 Conf: [[681 4] [7 50]]
5	10	20	Acc:0.98 Conf: [[690 7] [2 43]]	Acc:0.98 Conf: [[691 6] [6 39]]	Acc:0.92 Conf: [[670 27] [26 19]]	Acc:0.98 Conf: [[690 7] [1 44]]
6	10	25	Acc:0.98 Conf: [[701 4] [7 30]]	Acc:0.98 Conf: [[704 1] [8 29]]	Acc:0.93 Conf: [[684 21] [26 11]]	Acc:0.98 Conf: [[700 5] [7 30]]
7	10	30	Acc:0.98 Conf: [[700 6] [7 29]]	Acc:0.98 Conf: [[704 2] [11 25]]	Acc:0.94 Conf: [[688 18] [20 16]]	Acc:0.98 Conf: [[703 3] [8 28]]
8	10	35	Acc:0.98 Conf: [[696 6] [6 34]]	Acc:0.99 Conf: [[699 3] [4 36]]	Acc:0.93 Conf: [[672 30] [19 21]]	Acc:1.00 Conf: [[700 0] [0 42]]
9	1	1	Acc:0.98 Conf: [[697 3] [5 37]]	Acc:0.98 Conf: [[699 1] [12 30]]	Acc:.95 Conf: [[684 16] [19 23]]	Acc:0.98 Conf: [[699 1] [9 33]]
10	5	1	Acc:0.98 Conf: [[697 3] [6 36]]	Acc:0.98 Conf: [[697 3] [11 31]]	Acc:0.92 Conf: [[666 34] [23 19]]	Acc:0.98 Conf: [[699 1] [7 35]]
11	10	1	Acc:0.98 Conf: [[698 2] [6 36]]	Acc:0.98 Conf: [[698 2] [11 31]]	Acc:0.94 Conf: [[686 14] [28 14]]	Acc:0.99 Conf: [[700 0] [7 35]]
12	15	1	Acc:0.98 Conf: [[697 3] [5 37]]	Acc:0.98 Conf: [[697 3] [9 33]]	Acc:0.94 Conf: [[680 20] [24 18]]	Acc:0.98 Conf: [[699 1] [10 32]]
13	20	1	Acc:0.98 Conf: [[697 3] [6 36]]	Acc:0.98 Conf: [[697 3] [9 33]]	Acc:0.95 Conf: [[683 17] [19 23]]	Acc:0.98 Conf: [[700 0] [9 33]]
14	25	1	Acc:0.98 Conf: [[697 3] [5 37]]	Acc:0.98 Conf: [[697 3] [11 31]]	Acc:0.94 Conf: [[681 19] [20 22]]	Acc:0.98 Conf: [[699 1] [8 34]]
15	30	1	Acc:0.98 Conf: [[697 3] [6 36]]	Acc:0.98 Conf: [[697 3] [11 31]]	Acc:0.95 Conf: [[678 22] [11 31]]	Acc:0.98 Conf: [[698 2] [8 34]]
16	35	1	Acc:0.98 Conf: [[697 3] [6 36]]	Acc:0.98 Conf: [[698 2] [12 30]]	Acc:0.95 Conf: [[686 14] [16 26]]	Acc:0.98 Conf: [[699 1] [8 34]]

Accuracy is an evaluation of classifiers. Accuracy evaluates the number of correct prediction in the total number of predictions. In the machine learning, a 100% score means the best score and 0% error. Confusion matrix predicts the class or target variable, how much positive class or correctly classified and how much negative class or incorrectly classified and find negative outcomes or positive outcomes.

We have applied 16 numbers of iterations in all the observation for prediction, but we have found the highest accuracy in iterations 1, 8, and 11 with their corresponding confusion matrix. Iteration 1 has an accuracy of 99%, 97%, 94%, and 99% with seed value 1 and num-fold value 10 of classifiers decision tree, random forest, extra tree, and ensemble model, respectively. Iteration 8 has an of accuracy 98%, 99%, 93%, and 100% with seed value 35 and num-fold value 10 for all classifiers, respectively, as in iteration 1. Iteration 11 gives the accuracy of 98%, 98%, 94%, and 99% with seed value 1 and num-fold value 10 of all classifiers in the same respective way as in iterations 1 and 11. The various results obtained by previous studies are captured with our proposed study as shown in Table 3.

Table 3 Accuracy and techniques of old research details

Author	Techniques	Accuracy(%)
Prasad et al. (2016)	RST	99
	MST	99
Akbaş et al. (2013)	Bayes Net	98
	Naive Bayes	94
	SMO	94
	Ibk	91
	Random forest	99
Tyagi et al. (2018)	ANN	97
	KNN	98
	SVM	99
	DT	75
Ioniţă and Ioniţă (2016)	CART	89
	J48	89
	MLP	77
	RBF	79
	Naïve Bayes	70
Sivasakthivel and Shrivakshan (2017)	CART	86
Chaurasia et al. (2018a)	Naïve Bayes	97
	RBF	96
	J48	93
Verma et al. (1887)	CART	93
	SVM	92
	DT	94
	RF	94
	GBDT	95
	Ensemble	98
Verma et al. (2019)	PAC	97
	RNC	94
	BNB	96
	NB	95
	ETC	96
	Ensemble	99
Chang and Chen (2009)	NN	92
	DT	80
Cataloluk and Kesler (2012)	KNN	94
	Weighted K-NN	96

We have selected some old research work in the overall study during this research mentioned in Table 3. All the research work is related with thyroid and other medical data use in machine learning classifiers for prediction. The classification accuracy of this ensemble model is the highest (100%) compared with other classifiers mentioned worked in Table 3.

4 Conclusion

This research has been done on thyroid dataset by different machine learning classifiers such as decision tree, random forest tree, extra tree, and bagging ensemble model. The seed value 35 and num-fold value 10 have found the highest accuracy using bagging ensemble techniques. Therefore, bagging ensemble technique is the best compared with the other three classifier algorithms. In future work, we observe the identification of different affected factors of thyroid dataset and test more using different and large datasets for diabetes, heart disease, etc.

Acknowledgments The author is grateful to Veer Bahadur Singh Purvanchal University Jaunpur, Uttar Pradesh, for providing financial support to work as Postdoctoral Research Fellowship.

References

Akbaş A, Turhal U, Babur S, Avci C (2013) Performance improvement with combining multiple approaches to diagnosis of thyroid cancer. Engineering. 5(10):264–267

Alqurashi T, Wang W (2019) Clustering ensemble method. Int J Mach Learn Cybern 10(6):1227–1246

Aswathi AK, Antony A (2018) An intelligent system for thyroid disease classification and diagnosis. In: 2018 Second international conference on inventive communication and computational technologies (ICICCT). IEEE, pp 1261–1264

Azar AT, Hassanien AE, Kim TH (2012) Expert system based on neural-fuzzy rules for thyroid diseases diagnosis. In: Computer applications for bio-technology, multimedia, and Ubiquitous City. Springer, Berlin, Heidelberg, pp 94–105

Azar AT, El-Said SA, Hassanien AE (2013) Fuzzy and hard clustering analysis for thyroid disease. Comput Methods Prog Biomed 111(1):1–6

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Cataloluk H, Kesler M (2012) A diagnostic software tool for skin diseases with basic and weighted K-NN. In: 2012 International symposium on innovations in intelligent systems and applications. IEEE, pp 1-4

Chang CL, Chen CH (2009) Applying decision tree and neural network to increase quality of dermatologic diagnosis. Expert Syst Appl 36(2):4035–4041

Chaurasia V, Pal S, Tiwari BB (2018a) Chronic kidney disease: a predictive model using decision tree. Int J Eng Res Technol 11:1781–1794

Chaurasia V, Pal S, Tiwari BB (2018b) Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology 12(2):119–126

http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease 2013

Ioniță I, Ioniță L (2016) Prediction of thyroid disease using data mining techniques. BRAIN. Broad Research in Artificial Intelligence and Neuroscience 7(3):115–124

Landwehr N, Hall M, Frank E (2005) Logistic model trees. Mach Learn 59(1–2):161–205

Liu X, Xue A (2012) The thyroid disease analysis by a class of tissue P system. In: 2012 international symposium on information Technologies in Medicine and Education, vol 2. IEEE, pp 744–748

Ozyilmaz L, Yildirim T (2002) Diagnosis of thyroid disease using artificial neural network methods. In: Proceedings of the 9th international conference on neural information processing, 2002. ICONIP'02, vol 4. IEEE, pp 2033–2036

Prasad V, Rao TS, Babu MS (2016) Thyroid disease diagnosis via hybrid architecture composing rough data sets theory and machine learning algorithms. Soft Comput 20(3):1179–1189

Sharaff A, Gupta H (2019) Extra-tree classifier with metaheuristics approach for email classification. In: Bhatia S, Tiwari S, Mishra K, Trivedi M (eds) Advances in computer communication and computational sciences. Advances in intelligent systems and computing, vol 924. Springer, Singapore

Sivasakthivel A, Shrivakshan GT (2017) “A comparative study of diagnosing thyroid diseases using classification algorithm”. International Journals of Advanced Research in Computer Science and Software Engineering 7(Issue 8):ISSN: 2277-128X

Tyagi A, Mehra R, Saxena A (2018) Interactive thyroid disease prediction system using machine learning technique. In: 2018 Fifth international conference on parallel, distributed and grid computing (PDGC). IEEE, pp 689–693

Verma AK, Pal S, Kumar S (1887) Classification of skin disease using ensemble data mining techniques. Asian Pacific Journal of Cancer Prevention 20(6)

Verma AK, Pal S, Kumar S (2019) Prediction of skin disease using ensemble data mining techniques and feature selection method—a comparative study. Appl Biochem Biotechnol 27:1–9

Yadav DC, Pal S (2019) To generate an ensemble model for women thyroid prediction using data mining techniques. Asian Pac J Cancer Prev 20(4):1275–1281

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.