International Journal of Hospital Research 2022, 11(3)
http://ijhr.iums.ac.ir
Research Article

Open Access

A Model for Diagnosis of Thyroid Disease Based on Rules Extraction Using Tree Algorithms and Feature Selection

Maryam Talachian[1], Mohammad Fathian [1]

Information Technology Management, Iran University of Science and Technology, Tehran.Iran.

Abstract

Background and Objective: Proper and quick diagnosis of disease is necessary in the medical field for the correct and timely treatment. This issue becomes more important when faced to different diseases with similar symptoms, such as thyroid disease, which has similar symptoms to some disease such as cardiovascular disease. Data mining and machine learning techniques are reliable and valuable methods that can improve the ability of physicians for correctly diagnosis and treatment.

Method: The main goal of this research is to extract rules of thyroid disease, create the features and analyze feature selection algorithms including filter-based, wrapper based and the genetic algorithm to select the most effective features for thyroid diagnosis. The analysis also performed using decision trees models, random forest, bagging, boosting, and stacking methods for diagnosis and improvement of the illness classes precision that including Hypothyroidism and Hyperthyroidism. Model evaluation was performed with four metrics of accuracy, precision, recall, and F-measure.

Result and Conclusion: This research was conducted on data from the University of California (UCI), which included 7200 records with 21 features. Experimental results showed that the genetic algorithm (GA) has a maximum efficiency in feature selection, and the boosted tree with created features produced maximum F-measure among other classifier.

Keywords

Thyroid Disease, Data Mining, Tree Algorithms, Feature Selection

Background and Objective

Thyroid gland is one of the vital glands of the body affecting indirectly all of organs such as heart, kidney, and so on. This butterfly-shaped gland is located in the neck and front of the larynx. Thyroid releases two interconnected hormones called Thyroxin (T4) and Triiodothyronine (T3) that play a vital role in cell differentiation during embryo development. These hormones are important factors in making proteins, body temperature regulation, and energy generation in body[1] and are controlled by thyroid stimulating hormone, Thyrotropin (TSH). It is difficult to interpret results of in vivo experiments in the old, pregnant women, those who use medicines, and patients in specific medical conditions. Existence of large volumes of medical data leads to the production of data analysis tools for extracting hidden knowledge. Researchers have been focusing on application of data mining tools and statistical methods to improve large data analysis. The amount of data that needs to be processed goes to the big data category.

*Corresponding Author: Mohammad Fathian
Email: fathian@iust.ac.ir

Data mining is widely used to discover hidden information from large amounts of data. However, utilizing data mining in medical databases is a challenging process[2]. Data mining tools have published successful results in some scopes including disease diagnosis and prediction as well as recovery process. Since the last decade, Data mining is becoming increasingly important and tremendous amount of applications are underway in the healthcare industry, where most applications are introduced that can be classified into two categories: the bench of decision support and the development of policies[3].

In the field of "big data", recent developments in information and communication technologies are facilitating organizations to grow and innovate. Big data and smart healthcare systems independently attract much attention from both academia and industry. The combination of big data and smart systems can expedite the prospects of the healthcare industry[4].

Healthcare data have complexities and heterogeneity due to data nature. Data complexity and unavailability of large raw data and are some of the difficulties encountered[2]. It is difficult and impossible to discover the model of different kinds of data using traditional statistics; hence, data mining technics in medical data provides more effective analysis to improve diagnostic and care abilities. Using data mining in medicine is one of the practical technics of data mining, which playing a vital role in health and leading to discover a new, beneficial, and sustainable knowledge in database. Data mining and healthcare industry have emerged some of early detection systems and other healthcare related systems from the clinical and diagnosis data[5]. Disease diagnosis models classify disease diagnosis datasets to healthy people and patients using artificial intelligence and data mining approaches (such as neural network and rule-based approaches). New data mining approaches and classification methods have shown significant progress rather than statistical methods such as logistic regression. On the other hand, it is essential to use new technics in data mining such as hybrid methods like bagging, boosting, and stacking to solve classification models' problems such as non-linear relations between variables. One challenge in data mining algorithms is generating beneficent rules and outputs in field of disease diagnosis and coping with unbalanced data.

The problem of imbalanced class should be considered in using data mining techniques for predicting. A dataset is imbalanced if the classes are not approximately equally represented. There have been attempts to deal with imbalanced datasets in real issue such as pollution, risk management, fraud detection, and medical diagnosis[6]. in this case, if there is 1% patients and 99% healthy people, prediction process identifies all subjects as healthy if data are not balanced, while we look for high accuracy in diagnosis of number of the class with low population. Unbalanced data leads to high accuracy even if there is misdiagnosis and considering patients as healthy. Hence, it can be stated that total accuracy is not a suitable measure to evaluate unbalanced data. F-measure is efficient factor for unbalanced data.

This paper objectives are as follow:

1. Analysis of Hypothyroid, Hyperthyroid, using tree algorithms and ensemble methods based on critical factors as: total accuracy and F-measure
2. Creating new features from existing features of data to improve the accuracy of diagnosis
3. The comparison of some of the most famous methods of feature selection including filtering, wrapper, and GA methods and choosing effective methods also finding the best method for choosing the best feature in this data among the described methods
4. Generating beneficial rules of disease using tree algorithms

Related work

Many researchers have used data mining to diagnose diseases, Pavya and Srinivasan (2017) used filtering methods entitled F-Score, wrapper, and feature selection to diagnose diseases and classification, they used multilayer perceptron, back propagation neural network, support vector machine, and extreme machine learning. Results showed the highest efficiency of wrapper with accuracy of 98.14 among the algorithms[7]. Agrawal and Dhakar (2017) conducted feature selection to create a thyroid prediction system classifying it to four categories: overactive, underactive, patient and healthy; they used auto associative neural network (AANN) for modeling. Prediction accuracy was reported equal to 95.1[8].They used feature selection methods for their research.

Ahmad, Huang, and Shah (2017) proposed a hybrid diagnostic system for thyroid disease diagnosis. Data output was used as the input of fuzzy systems and accuracy of classification reached 99.1[9]. Pan and Zhang (2016) introduced a new method for classification based on random forest and claimed that the accuracy of this method is higher than bagging, boosting, and random forest methods; they obtained accuracy of 95.63 for UCI data and of 96.16 for the clinical data. That study applied decision tree 4.5c as the base model and used random forest algorithms to classify data with optimal and low indicators[10]. Banu (2016) used decision tree and J48 algorithms to diagnose Hypothyroidism through Weka software; they obtained accuracy rates of 95.38 and 99.57 using decision tree and J48 methods, respectively[11]. Sangeetha and Planivel (2018) used Naive Bayse, random forest, random tree, and D tree through Weka software to anticipate thyroid disease; they also applied crossover validation with k-folds and k=6 then compared results. The highest accuracy (99.80) was related to decision tree algorithm[12].

The following researches apply UCI [1] database of California University related to thyroid disease that we use in this article. Pavya and Srinivasan (2018) used ant colony (ACO), k-nearest neighbor (KNN), wrapper, and support vector machine and reported total accuracy equal to 94[13]. Umadevi and Jeen Marseline (2017) applied KNN and neural network and reported the highest accuracy equal to 90[14]. Pal, Anand, and Kumar Dubey (2017) used three data mining techniques of Naive Bayse, KNN., and support vector machine by Weka Software; which KNN among them had the highest reported accuracy of 96.90[15]. Chandel and Kunwar (2016) applied Naive Bayse, KNN, and support vector machine and the highest accuracy reported was 93.44[16]. Kabir and Shahjahan (2012) reported accuracy in thyroid disease to 99 using ant colony and a hybrid method[17]. Seyti and Aliari (2008) used GA and neural network and improved accuracy of training data up to 99%[18]. These group of researchers used KNN, Naïve Bayes, support vector machine, Ant colony and neural network. Their highest accuracy in this data was 99 and their critical factor was only accuracy. In this studies, generating rules and using other criteria factors such as f-measure for results not reported.

Method

This paper conducted on UCI thyroid data base with three classes of disease including: healthy, hypothyroid patients, and hyperthyroid patients. This research is based on the CRISP process including steps as follows: Disease understanding, Data understanding, Data preparation, Modelling and Evaluation. And 10-fold cross validation method was used for validation. The evaluation is done with the criteria of total accuracy, F-measure, Recall and precision. We try to select the best model among the data mining models and improve the accuracy and other criteria for thyroid diagnosis. Therefore, we use feature selection and feature creation and the results of the models mentioned in this study to introduce the most efficient model for disease detection on the data of this research. This study was performed using Rapid Miner software.

The proposed method is shown in **Figure 1**.

[1] https://archive.ics.uci.edu/ml/datasets.html

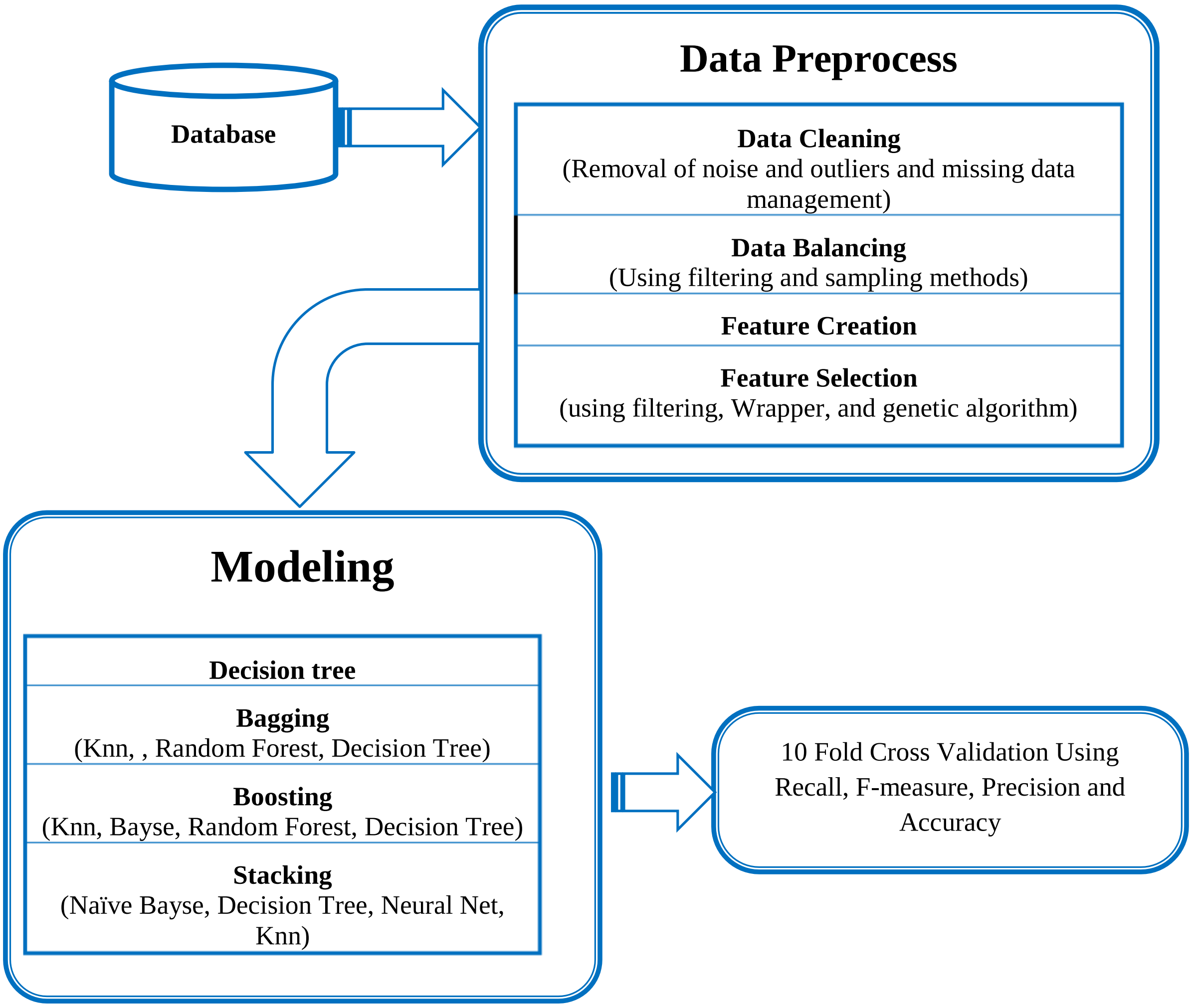
Database
Data Preprocess
Data Cleaning
(Removal of noise and outliers and missing data management)
Data Balancing
(Using filtering and sampling methods)
Feature Creation
Feature Selection
(using filtering, Wrapper, and genetic algorithm)
Modeling
Decision tree
Bagging
(Knn, , Random Forest, Decision Tree)
Boosting
(Knn, Bayse, Random Forest, Decision Tree)
Stacking
(Naïve Bayse, Decision Tree, Neural Net, Knn)
10 Fold Cross Validation Using Recall, F-measure, Precision and Accuracy

Figure 1. Proposed method

Data

The required data of paper were extracted from UCI database of California University including 7200 thyroid disease and 21 features which includes the main features of thyroid disease such as T3, T4, TSH, FTI, T4U. This data contains three classes of overactive, underactive, and healthy category; 21 new features were added to this dataset in order to improve final precision of prediction based on the recommendation of an expert physician and mathematical-statistical operations on numerical features. Preparation process was done on main data and data with created features through Rapid Miner software. Some methods were applied on datasets within different preparation steps.

Pre-processing

Data preparation methods in this study includes removal the duplicated data, handling outliers and missing data, data balancing, discretization, and feature selection.

Balancing

These data consist of three classes of healthy, underactive, and overactive with lower number of data in disease class compared to healthy class data. Since machine-learning algorithms are planned for populated class prediction, data balancing would be a challenge to have high precision of disease classes. It should be noted that balancing process was implemented only on training data to prevent false increasing precision. **Figures 2** and **3** show data charts before and after balancing. As shown in **Fig**. **2**, the horizontal axis represents the classes of disease that are class 3 healthy class 2 hyperthyroid and class 1 hypothyroidism.

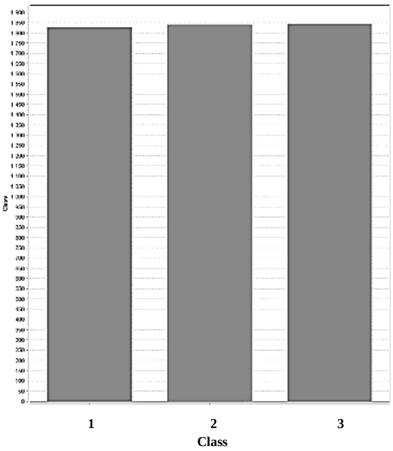

Figure 2. Data charts before balancing

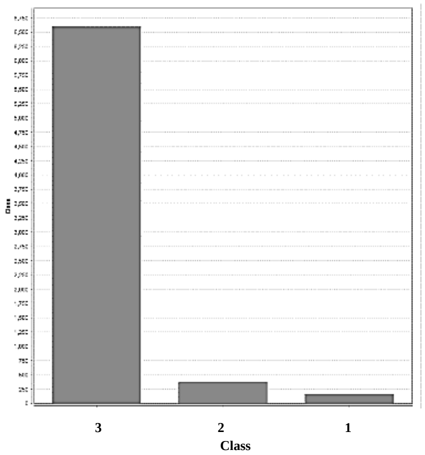

Figure 3. Data charts after balancing

Feature Selection

Feature selection allows us to reduce computation time and have a better performance in prediction; in this regard, we will have better understanding of data and patterns in machine learning. Feature selection method focuses on selecting a set of input variables that have the highest effect on input data, and on reducing noise and leads to a good anticipation. Some of feature selection approaches include filtering, wrapper, use of classifiers such as support vector machine and GA[19] Filtering, wrapper, and Genetic Algorithm approaches were used in this research, as they are described in this section.

Filtering

Filtering-based approaches include three main steps;

A subset of the whole set features is selected, measured, then tested by machine learning algorithms. So initially, a subset of features is created then its data is measured and this process repeated until the result is matched with stopping criterion. Stopping criterion can be a threshold of measurement results. If the result has not reached to the threshold, a new subset is created and the process is repeated. Hence, the final subset has some features with more information. At the last step, means experiment step, those features are tested by machine algorithms. Filtering-based approaches use statistical tests for determining a subset of features with the highest precision in prediction[7]. In this article we use 2 methods of filtering, weighted by Information gain and Gini index. In these methods, the attributes are weighted by the statistical methods of information gain and Gini index. The most influential attribute receives the most value and the least effective receives the least value. Gini Index is an impurity splitting method. It is suitable to binary, continuous numeric type values. It was proposed by Breiman in 1984 and has widely been used in algorithms[20]. The information gain measure is used to select the test attribute at each node of the decision tree. The information gain measure prefers to select attributes having a large number of values[21].

Wrapper

Like filtering method, a subset of features is selected in wrapper method but measurement step is done using machine learning algorithms and wrapper methods are time consuming. Wrapper, on the other hand, uses machine-learning algorithms for feature selection; hence, it leads to better results than filtering methods[7].

Genetic Algorithm (GA)

GA is a random search method based on the evolutionary biological process whose important feature is the ability to execute it in large search spaces without getting stuck in the local extreme. Therefore, due to the possibility of several local optimizations, this algorithm can be considered as a suitable method for selecting features. GA selects a subset randomly from possible solutions within optimization process, then features are created in a new subset considering properties of each feature using genetic performances like crossover and mutation until the chosen criterion is prepared[22]. We used GA to optimize selection evolutionary in Rapid Miner software and adjusted its suitable parameters. The best stop criterion is determined by the software. The GA method provides better results than the other methods, which is why it was used in this study. Results of feature selection methods have been compared in Table 1.

Table 1. Comparison of feature selection results

Feature selection method	Total accuracy	F-measure	Number of selected features	Effective Features
Filtering: Weight by information gain	99.30	97.18	10	Class, age, sex, query-hypothyroid, psych, TSH, on-thyroxin, T3, TT4, T4U, FTI
Filtering: Weight by Gini index	99.18	97.13	10	Class, age, sex, query-hypothyroid, psych, TSH, on-thyroxin, T3 TT4, T4U, FTI,
Wrapper forward selection	99.54	97.53	7	Thyroid surgery, on-thyroxin, TSH, T3, TT4, T4U, FTI
Feature selection with GA	99.63	98.16	13	on-thyroxin, thyroid surgery, TSH, T3, TT4, T4U, FTI, sick, treatment, antithyroid medication, goiter, tumor, query- hypothyroid

Feature Creation

indicates these features after selecting features using GA[2].

To improve prediction precision, 21 new features was proposed by the expert physician and added to the features through performing mathematical and statistical operations on numerical properties and method combining existing features; total features reached to 42. New features have been created by division of the numerical effective features into their means as well as their differences with the mean of the same features and also minus the features from each other and dividing the features into each other. **Figure 2**

[2] Genetic Algorithm

Table 2. Effective features after feature creation

Feature selection method	Accuracy	F-measure	Number of selected features	Effective features after considering new created features
Selecting features using GA	99.66	98.33	17	T3, TSH, on-thyroxin, thyroid- Surgery, Query-Hypothyroid, Goiter, FTI, Antithyroid Medication, TT4 subTSH, T3 subTSH, T3byFTI, T3subFTI, T3subTSH, Pregnant, Sick, Psych, TT4byT4U

There is a commonality between the features selected in Tables 1 and 2. As mentioned after comparing feature selection results, feature selection method using GA had better performance on our data in both data with created features and with original features. There are 9 features that are common among both of them, features are as follow: T3, TSH, on-thyroxin, thyroid- Surgery, Query-Hypothyroid, Goiter, FTI, Anti-thyroid Medication, Sick and 2 features in the original data that are not common among these two tables: Psych, pregnant. In table 2, 6 features were selected from created features. It should be noted that all features are characteristics of thyroid disease.

Modelling

Classification

Classification divides sample data to target classes and classification techniques predicts target class for each data. For instance, a patient can be classified to a patient with low-risk or a high-risk patient based on the disease pattern and patient's data. This classification method is recognized as a supervised learning using target classes. Binary and multilevel are two classification methods that target classes in binary method includes two classes such as high-risk and low-risk, while there are more than two classes in multilevel method. In this method, data are divided into two parts of train and test data. Learning is done using train data and prediction accuracy and precision of classifier is done by test data. Classification is one of the most widely used method in the field of medicine and health. Classification techniques are used to predict disease process[23]

Decision Tree

Decision tree was proposed in 1993. The base model of decision tree consists of a root node, some internal nodes, and a set of terminal nodes. Data are classified based on the classification pattern in decision. There must be a rule for each node. Decision trees are popular for their high accuracy in results compared with other methods[24]. Decision tree may be binary or multilevel. If decision tree is binary, each node will have two children but number of elements is key point in non-binary types. The tree is classified in accordance to expected values given to each branch. The main key for decision tree creation is selection of the best feature in node division. Information benefits, Gini Index, and so on are methods used to find the best feature[25]. Decision tree is like a flowchart in which those nodes that do not have leaf determine a test on a specific feature, each branch determines test result, and each node has a class label. The highest node with the most labels is the root node. Decision tree is a classifier that uses tree-like graph[26]. This study used different types of decision trees like C4.5, ID3, and random forest for modelling and prediction and disease rules were extracted.

Combined Algorithms

Traditionally, boosting is a general method to combine rules using several weak classifiers to create a hybrid classifier with high accuracy.

Boosting algorithms is known as PAC (Probably Approximately Correct) in machine learning. In this method, weak learning algorithms, which are a little better than random guessing, change to a strong learning algorithm with high accuracy[27]

Stacked generalization that is called stacked regression is a hybrid method used to combine different models; in brief, it can be stated that this model combines different models to create a new model for prediction so that this model provides a better prediction performance compared to its components. Hybrid models reduce errors potentially. However, hybrid models are based on train data without considering model complexity and its generalization to new data. To improve this situation, weight mean model is used; this model gives the heavier weight to the model that is highly matched with data[28].

If combined learning is used for prediction, it will be called Bagging. Bagging aimed at mainly to make predictors more independent and improve prediction accuracy using majority vote. Bagging uses sampling method with replacement and one sample may be selected several times or not be selected at all. These frequent samplings increase independence among predictors in a random way. Bagging creates bootstrap aggregation which creating required diversity by repeating different samplings. The considered class will be determined by entering a new sample considering spatial neighbors and the collective vote[29]. In this paper, combined algorithms of bagging and boosting and stacking with decision tree, neural network, KNN, and Naive bayse models used to predict thyroid disease.

Criteria factors

Accuracy is the most important criterion for determining the efficiency of a classification algorithm. This criterion calculates the total accuracy of a classifier. In fact, this criterion shows that a few percent of our entire set of experimental records are correctly classified. The accuracy of a model is calculated in **Formula 1**:

Formula 1. Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where

TP is True Positive

TN is True Negative

FP is False Positive

FN is False Negative

Other important criteria especially for imbalanced data are F-measure, recall and precision as shown in **Formula 2, 3, 4** respectively**.**

Formula 2. Precision

$$precision = \frac{\sum true\ positive}{\sum test\ outcome\ positive}$$

Formula 3. Recall

$$recall = \frac{\sum true\ positive}{\sum condition\ positive}$$

Formula 4. F-measure

$$F - measure = \frac{precision \times recall}{precision + recall} * 2$$

Results and Evaluation

All parameters of each model were prepared and optimized several times to achieve the best result. Parameters were optimized by GA separately and within the combined models. The best output of each model is indicated in the tables. 10-fold cross validation was used to validate model and test it, then, results were compared using total accuracy and f-measure criteria.

Disease Rules

Comparing results of decision trees, decision tree C4.5 showed the best results among other decision trees, and disease rules were extracted from this tree. These rules were extracted with both main and created features of data for

Hypothyroidism and Hyperthyroidism classes. Hypothyroidism rules, Hyperthyroidism rules, and created features are shown in **Table 3** & **Table 4**. Moreover, Hyperthyroidism rules with and without features are indicated in **Table 5** & **Table 6**.

Table 3. Hypothyroidism rules

If	FTI	TSH	On_thyroxine	TT4	Thyroid _surgery	TT4	T3	samples	confidence
1	≤ 0.064	> 0.006	-	-	-	-	≤ 0.026	1830	99.78%

Table 4. Hypothyroidism rules with created features

If	TT4 Sub TSH	TSH	FTI	T3	T3 by FTI	T3 Sub TSH	T3 Sub FTI	On_Anti thyroid_ medication	TT4 Sub TSH	T3 Sub TSH	T3 Sub TSH	TSH	Sample	Conf
1	> -0.00	> 0.00	≤ 0.06	-	≤ 0.56	≤ 0.01	-	0	≤ 0.06	> 0.00	-	-	4	75%
2	> -0.00	> 0.00	≤ 0.06	-	≤ 0.56	≤ 0.01	-	0	≤ 0.06	≤ 0.00	> -0.0	-	47	100%
3	> -0.00	> 0.00	≤ 0.06	-	≤ 0.56	≤ 0.01	-	0	≤ 0.06	≤ 0.00	≤ -0.0	> 0.04	4	100%
4	≤ -0.00	-	≤ 0.06	≤ 0.03	-	-	> -0.05	-	-	-	-	-	111	97.29 %

Table 5. Hyperthyroidism rules

If	FTI	TSH	On_thyroxine	TT4	Thyroid _surgery	TT4	T3	Samples	Confidence
1	> 0.064	> 0.006	0	≤ 0.153	0	> 0.054	-	1831	99.95%
2	> 0.064	> 0.006	0	≤ 0.153	0	≤ 0.054	≤ 0.016	10	100%

Table 6. Hyperthyroidism rules with created features

If	TT4Sub TSH	TSH	FTI	T3	T3	Thyroid surgery	T3	On_thyroxin	TT4Sub TSH	TT4Sub TSH	samples	conf
1	> -0.00	> 0.006	> 0.064	≤ 0.043	> 0.029	0	-	0	≤ 0.14	> 0.10	9	88%
2	> -0.00	> 0.006	≤ 0.064	≤ 0.043	≤ 0.029	0	> 0.003	0	≤ 0.14	-	364	98.3%

Discussion

The represented results of all models created in this research using recall, precision, F-measure and total accuracy are indicated in **Table 7**.

Table 7. Results of models

Method	**Total accuracy**	**Accuracy of Hyperthyroidism class**	**Accuracy of Hypothyroidism class**	**Recall**	**Precision**	**F-measure**
Decision tree	99.41	99.73	100	99.70	95.11	97.35
Boosting decision tree	99.63	98.64	98.19	98.86	97.47	98.16
Bagging decision tree	99.59	98.91	95.78	98.14	97.19	97.66
Decision tree with created features	99.61	99.46	97.59	98.90	97.0	97.94
Boosting decision tree with created features	99.66	99.46	97.59	98.92	97.71	98.31
Bagging decision tree	99.59	99.46	98.19	99.10	96.63	97.84
Id3	96.16	88.32	81.33	88.97	76.26	82.13
Boosting Id3	96.67	84.78	72.89	85.28	81.21	83.19
Bagging Id3	95.24	70.11	36.14	68.17	75.52	71.65
Boosting KNN	96.59	54.35	77.71	77.20	91.83	83.88
Bagging KNN	95.88	36.68	73.49	70.02	95.25	80.71
Boosting Bayse	80.93	83.70	93.98	86.05	64.02	73.41
Bagging Bayse	95.26	29.35	84.94	71.19	86.38	78.05
Random forest	99.30	100	100	99.77	94.55	97.08
Boosting random forest	99.25	92.39	96.39	96.19	97.04	96.61
Bagging random forest	96.14	35.33	90.96	75.36	79.29	77.27
Stacking	99.41	99.73	100	99.70	95.11	97.35
Stacking with created features	99.59	98.91	97.59	98.72	97.18	97.94

According to comparison of total results with total accuracy and F-measure in **Table 7**, decision tree models are the best models. Although some models have similar accuracies, F-measure (as the most efficient criterion for infrequent classes) was the most valuable criterion to consider prior models.

Results

We compare results obtained from models to predict thyroid disease on the thyroid UCI database that is used in this paper. As explained before, preparation and balancing were performed on the data and then feature selection was used to improve the accuracy of thyroid diagnosis. The results of feature selection were presented in **Table** 1. New features were created on the data that mentioned in the feature creation section, and again the features were reselected with these new features, the results were presented in **Table 2**.

Results of these models are indicated in Table 8. As explained above, in this paper, new features were created with the initial features of the data. Therefore, in order to compare the prediction efficiency of the features created, with original features of data, re-modelling was performed on the models with the best predictive outcomes, as shown in Table 8. The comparison of the results of the models with the original features and models with created features is presented in **Table 9** with the criteria of total accuracy and f-measure. Comparison of results of best models for research data with and without features using accuracy and f-measure criteria is shown in **Table 9**.

Table 8. Results of best models (without feature creation)

Method	Total accuracy	Accuracy of Hyperthyroidism class	Accuracy of Hypothyroidism class	Recall	Precision	F-Measure
Decision tree	99.41	99.73	100	99.70	95.11	97.35
Boosting decision tree	99.63	98.64	98.19	98.86	97.47	98.16
Bagging decision tree	99.59	98.91	95.78	98.14	97.19	97.66
Stacking	99.41	99.73	100	99.70	95.11	97.35

Table 9. Comparison of results of total accuracy and f-measure with and without feature creation in best models

Method	Total accuracy	Total accuracy with feature creation	f-measure	f-measure with feature creation
Decision tree	99.41	99.61	97.35	97.94
Boosting decision tree	99.63	99.66	98.16	98.31
Bagging decision tree	99.59	99.59	97.66	97.84
Stacking	99.41	99.59	97.35	97.94

As you see in **Table 9** created features have improved total accuracy and f-measure of best models in **Table 8**. Although total accuracy has not changed with created features in bagging method, the more valuable metric (f-measure) have improved. Hence, table 9 indicates that the highest accuracy and f-measure is related to boosting decision tree with created features.

The results of past research on this data show that Pavya and Srinivasan (2018)[13] reported accuracy of equal 94%, Umadevi and Jeen Marseline (2017)[14] the highest accuracy of 90% , Pal, Anand, and Kumar Dubey (2017)[15] accuracy of 96.90%, Chandel and Kunwar (2016)[16] accuracy of 93.44%, Kabir and Shahjahan (2012)[17] accuracy of 99% and finally Seyti and Aliari (2008) accuracy of 99%[18]. Therfore, their highest accuracy on this data was 99%. In this study, the total accuracy improved to 99.63% and F-measure to 98.31, and other criteria that are appropriate for this type of data, which have not been addressed in previous researches, have also been reported.

Conclusion

In this research we focus on diagnosis of thyroid disease. This study was performed on data from three categories of people with hyperthyroidism and hypothyroidism and healthy ones to predict health, underactive and overactive thyroid classes. New features have been created combining existed data features and improved prediction accuracy and F-measure. This improved the results of previous researches. Filtering, wrapper, and GA methods were analysed to select features. GA was introduced as the most effective method for feature selection on this data. Decision tree, C4.5, Id3, random forest, bagging, boosting, and stacking models was conducted on the data. Disease rules were extracted. It should be noted that this paper assessed 4 metrics including total accuracy, f-measure, recall, and precision at each step.

While data mining methods can supply detailed insight into health models, this approach has notable limitations. Given the small number of health centres that provide patient data for research, our results are not generalizable. For example, we tried to obtain expert physicians' approval of the results and rules of the disease extracted from this study but this was not possible due to the physicians' emphasis on conducting clinical examinations on individual patients and physically lack of access to the patients population (there was only access to patients' symptoms and laboratory results).

Using hybrid methods and integrating them with deep learning to enhance the accuracy of prediction in data mining algorithms can be very useful. We can also apply other techniques such as image processing and deep learning to diagnose thyroid cancer by using patient scanning, and especially where we have inaccuracy for experts to diagnosis or the error in diagnosis is high.

Perhaps the biggest challenge of data mining in the health and medical field is the lack of access to data in this area. Patients' data are either unrecorded or too scattered in medical records, and are usually manually recorded without any particular order. Since these data are typically large in volume, extracting and ordering them requires a very long time by specialist physicians, which makes these data often unusable for data mining. Using patient electronic records or a very simple Excel file in each treatment center and entering the symptoms, several influential patient features, without the need for a specialist and with little time for each client, can help to access this valuable data.

Conflict of interests

None.

Authors' contributions

The authors are the same

References

1- Kodaz, H., Özşen, S., Arslan, A., & Güneş, S. (2009). Medical application of information gain based artificial immune recognition system (AIRS): Diagnosis of thyroid disease. Expert Systems with Applications, 36(2), 3086-3092.

2- Tan, J., Xiong, T., Miao, H., Sun, R., & Wu, M. (2018, April). A case study of medical big data processing: Data mining for the hyperuricemia. In 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA) (pp. 196-201). IEEE.

3- Sohail, M. N., Jiadong, R., Uba, M. M., & Irshad, M. (2019). A Comprehensive Looks at Data Mining Techniques Contributing to Medical Data Growth: A Survey of Researcher Reviews. In Recent Developments in Intelligent Computing, Communication and Devices (pp. 21-26). Springer, Singapore.

4- Pramanik, M. I., Lau, R. Y., Demirkan, H., & Azad, M. A. K. (2017). Smart health: Big data enabled health paradigm within smart cities. Expert Systems with Applications, 87, 370-383.

5- Jothi, N., & Husain, W. (2015). Data mining in healthcare–a review. Procedia Computer Science, 72, 306-313.

6- Sohrabi, M. K., & Akbari, S. (2016). A comprehensive study on the effects of using data mining techniques to predict tie strength. Computers in Human behaviour, 60, 534-541.

7- Pavya, K., & Srinivasan, B. (2017, March). Feature selection algorithms to improve thyroid disease diagnosis. In Innovations in Green Energy and Healthcare Technologies (IGEHT), 2017 International Conference on (pp. 1-5). IEEE.

8- Agrawal, K., & Dhakar, M. (2017). Thyroid Prediction System using Auto Associative Neural Network. Thyroid, 6(4).

9- Ahmad, W., Huang, L., Ahmad, A., Shah, F., & Iqbal, A. (2017). Thyroid diseases forecasting using a hybrid decision support system based on ANFIS, k-NN and information gain method. J. Appl. Environ. Biol. Sci, 7(10), 78-85.

10- Pan, Q., Zhang, Y., Zuo, M., Xiang, L., & Chen, D. (2016, December). Improved Ensemble Classification Method of Thyroid Disease Based on Random Forest. In Information Technology in Medicine and Education (ITME), 2016 8th International Conference on (pp. 567-571). IEEE.

11- Banu, G. R. (2016). A Role of decision Tree classification data Mining Technique in Diagnosing Thyroid disease. International Journal of Computer Sciences and Engineering, 4(11), 111-115.

12- Sangeetha, S., & Palanivel, K. (2018). Anticipating Thyroid Disorders using Data Mining Techniques.

13- Pavya, K., & Srinivasan, B. (2018). Enhancing Wrapper Based Algorithms for Selecting Optimal Features from Thyroid Disease Dataset.

14- Umadevi, S., & Jeen Marseline, K. S. (2017). Applying Classification Algorithms to Predict Thyroid Disease. International Journal of Engineering

15- Pal, R., Anand, T., & Dubey, S. K. (2018). Evaluation and performance analysis of classification techniques for thyroid detection. International Journal of Business Information Systems, 28(2), 163-177.

16- Chandel, K., Kunwar, V., Sabitha, S., Choudhury, T., & Mukherjee, S. (2016). A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques. CSI transactions on ICT, 4(2-4), 313-319.

17- Kabir, M. M., Shahjahan, M., & Murase, K. (2012). A new hybrid ant colony optimization algorithm for feature selection. Expert Systems with Applications, 39(3), 3747-3763.

18- Aliari, F. S. M. (2009). Diagnosis of thyroid disease using neural network and genetic algorithm, The 2nd Joint Congress on Fuzzy and Intelligent Systems of Iran, Sivilica.

19- Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28.

20- Manek, A. S., Shenoy, P. D., Mohan, M. C., & Venugopal, K. R. (2017). Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. World wide web, 20(2), 135-154.

21- Bhatt, S., Cameron, E., Flaxman, S. R., Weiss, D. J., Smith, D. L., & Gething, P. W. (2017). Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalization. Journal of The Royal Society Interface, 14(134), 20170520.

22- Suzuki, T., & Ohkura, Y. (2016). Financial technical indicator based on chaotic bagging predictors for adaptive stock selection in Japanese and American markets. Physica A: Statistical Mechanics and its Applications, 442, 50-66.

23- Otukei, J. R., & Blaschke, T. (2010). Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms. International Journal of Applied Earth Observation and Geoinformation, 12, S27-S31.

24- Reyhani Nia, F. M. B. B. R. (2015). Reduction of classification error in determining the thyroid disease in Shoushtar city using a tree boosting algorithm, Journal of North Khorasan University of Medical Sciences 7 (2): 183-391.

25- Otukei, J. R., & Blaschke, T. (2010). Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms. International Journal of Applied Earth Observation and Geoinformation, 12, S27-S31.

26- Rutkowski, L., Jaworski, M., Pietruczuk, L., & Duda, P. (2014). Decision trees for mining data streams based on the Gaussian approximation. IEEE Transactions on Knowledge and Data Engineering, 26(1), 108-119.

27- Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. International Journal of Bio-Science and Bio-Technology, 5(5), 241-266.

28- Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. International Journal of Information Technology and Knowledge Management, 2(2), 271-277.

29- Ommati, M., Sahebi. (2016). Detection of changes in SAR polar metric images based on improved water-absorption algorithm, Journal of Science and Technology of Mapping, 6 (2), 63-78.

Please cite this article as:

Maryam Talachian, Mohammad Fathian. A Model for Diagnosis of Thyroid Disease Based on Rules Extraction Using Tree Algorithms and Feature Selection. Int J Hosp Res. 2022; 11 (3).