Prediction Of Thyroid Disorders Using Advanced Machine Learning Techniques

Priyanka Duggal
Department of CSE
Amity University Uttar Pradesh,
Noida, UP, India
Email: priyankaduggal05@gmail.com

Shipra Shukla
Department of CSE
Amity University Uttar Pradesh,
Noida, UP, India
Email: ershiprashukla88@gmail.com

Abstract**—** The paper presents several methods of feature selection and classification for thyroid disease diagnosis, related to the machine learning classification problems. Two common diseases of the thyroid gland, which releases thyroid hormones for regulating the rate of body's metabolism, are hyperthyroidism and hypothyroidism. Classification of these thyroid diseases is a considerable task. An important problem of pattern recognition is to extract or select feature set, which is included in the pre-processing stage. The proposed methods of feature selection are Univariate Selection, Recursive Feature Elimination and Tree Based Feature Selection. Three classification techniques have been used namely Naïve Bayes, Support vector machines and Random Forest. Results shows that the Support Vector Machines are the most accurate technique and hence this was used as a classifier to separate the symptoms of thyroid diseases into 4 classes namely Hypothyroid, Hyperthyroid, Sick Euthyroid and Euthyroid (negative).

Keywords**—***Recursive Feature Elimination, Univariate Selection, Tree Based technique, Naïve Bayes, Support Vector Machines, Random Forest, Hypothyroid, Hyperthyroid, Euthyroid*

I. INTRODUCTION

Thyroid diseases are increasing in magnitude everyday and spreading all over the world. The thyroid is a gland that produces thyroid hormone. One out of ten Indians suffer from Thyroid disorders. This disorder primarily takes place at between the age of 17-54. Identifying the thyroid disorder from the laboratory test report is very complex and requires extensive knowledge and experience , hence using machine learning techniques for this purpose makes the task easier and the results more accurate. The Thyroid dataset can be processed and after actual usage provide important facts and information for decision –making and diagnosing the disorders faster and more accurately and hence increasing the survival chances of patients . This hormone regulates vital body functions like breathing, body weight, heart rate, muscle strength. . This hormone regulates important body functions like weight, heart rate, muscle health and breathing. Studies state that one out of ten Indians suffer from a thyroid disorder. Thyroid disorders cause an increase in blood sugar, cholesterol level, depression, low fertility levels and cardiovascular, complications. The important factors that are majorly responsible for the abnormal function of the thyroid gland and the improper secretion of thyroid hormones are infection, trauma and stress.

The patients shall be classified into the following **4 classes** of thyroid disorders:

i. **Hypothyroid-** The thyroid gland fails to make sufficient amount of hormones and results in slowing down many of the body's functions. The Symptoms include Increased TSH , Decreased FT4, Weight Gain, Decreased Appetite, Slow Pulse and Fatigue, Decreased Metabolism

ii. **Hyperthyroid-** The thyroid gland makes excess of hormones than the body requires, which results in speeding of the body's functions. The Symptoms include Decreased TSH, Increased FT4, Weight Loss, Increased Appetite, Increased Pulse, Sweating and Increased Metabolism.

iii. **Sick Euthyroid-** The level of hormones made by the thyroid gland are comparatively lower in euthyroid(normal functioning thyroid gland) patients with temporary illness. The Symptoms include Increased FT4 and Increased FT3.

iv. **Euthyroid (negative)-** A normal functioning thyroid gland.

The performance measure is calculated from the confusion matrix with the accuracy. Experimental results were obtained from the WEKA (Waikato Environment for Knowledge Analysis).

II. RELATED WORK

Prerana ,Parveen Sehgal and Khushboo Taneja, [1] proposed a technique for detecting the thyroid by utilizing the back propagation algorithm. ANN(Artificial Neural Networks) is developed using the back propagation of error to identify preliminary thyroid prediction , it is trained using different training datasets . MATLAB was used to provide experimental results.

Ling Chen, Xue Li, Quan Z. Sheng and Wen-Chih Peng [2] proposed an expert system comprising of 3 stages using the support vector machines model for thyroid disease diagnosis.

Ammulu & Venugopal [3] utilized the random forest approach to predict the hypothyroid disorder by collecting the dataset

978-1-7281-2791-0/20/$31.00 ©2020 IEEE

from UCI repository. The performance measure is calculated from the confusion matrix with the accuracy

Shankar and Lakshman [4] proposed a multi-kernel support vector machine model and optimal feature selection to classify thyroid patients.

Irina Ionita [5] proposed a comparative study between various classification models like multi- layer perceptron, Naïve Bayes , Radial Basis Function Network and Decision Tree on the Thyroid Disease Dataset.

Kulkarni and Karwankar [6] proposed a MFHLSCNN (Modified Fuzzy Hyper Line Segment Clustering Neural Network) algorithm to classify thyroid patients.

Vinod and Vimal [7] proposed the idea of an intelligent system to detect thyroid diseases in pregnant ladies through Artificial Neural Networks.

Ahmed and Soomrani [8] proposed a TDTD framework (Thyroid Disease Type Diagnostics) which helps in cleaning of medical data and helps physicians in diagnosing thyroid disorders.

Pandey Tiwari, A. Shrivas, and A. K. Sharma, [9] proposed an ensemble model with feature selection to classify thyroid patients.

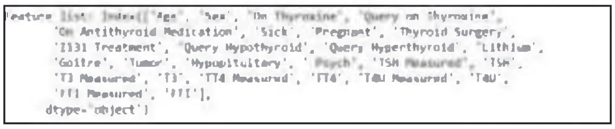
Termutas [10] proposed a comparative study of various machine learning models and used neural networks for classification of thyroid disorders.

III. PROBLEM DESCRIPTION & JUSTIFICATION

More than half of the Indian population suffers from undiagnosed or misdiagnosed thyroid diseases. Women are seven times more likely to contract thyroid problems than men and nearly half of all women and a quarter of all men in India will die with evidence of an inflamed thyroid. The symptoms of this disease often vary from person to person and are non-specific, so a correct diagnosis can easily be missed or misdiagnosed for irrelevant issues. Finding an accurate solution to this problem for healthcare practitioners via Classification techniques for diagnosing/classifying a particular thyroid disease that a person may have will cause an immense decrease in misdiagnoses as it is capable of distinguishing between problems of the thyroid gland and other illnesses in the body as well as providing the ability to detect the disease before it forms into a more destructive anomaly.

A. Dataset Description

The Dataset was extracted from The UCI Machine Learning Repository. This dataset was used for research, development and experimental purposes. The number of instances is 7200 instances and 27 attributes. The dataset contains the following attributes:



```
Feature list: Index(['Age', 'Sex', 'On Thyroxine', 'Query on Thyroxine',
       'On Antithyroid Medication', 'Sick', 'Pregnant', 'Thyroid Surgery',
       'I131 Treatment', 'Query Hypothyroid', 'Query Hyperthyroid', 'Lithium',
       'Goitre', 'Tumor', 'Hypopituitary', ' Psych', 'TSH Measured', 'TSH',
       'T3 Measured', 'T3', 'TT4 Measured', 'TT4', 'T4U Measured', 'T4U',
       'FTI Measured', 'FTI'],
      dtype='object')
```

Fig. 1. Set of attributes of the database

IV. FEATURE SELECTION

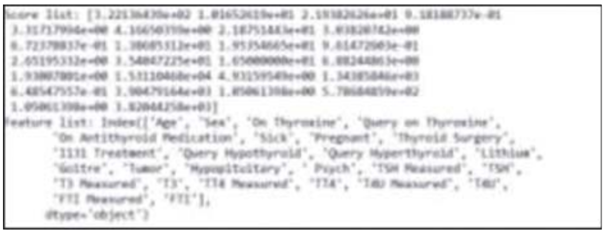
Feature Selection the most essential task in every Machine Learning model and it involves selecting the most important and related features that affect the accuracy of determining the target variable and discarding the features that are not so important. It is important to get rid of the insignificant features because they can result in decreasing the accuracy of the model. Some advantages of feature selection are as follows:

a. **Reduction in Overfitting**- It makes the data less redundant hence reducing the possibility of making decisions based on irrelevant features.
b. **Improvement in Accuracy**- The data becomes refined which makes it less misleading which increases the model accuracy.
c. **Reduction in Training Time** –The data points become less in number which helps in reducing the algorithm complexity and helps the algorithms to train faster.

Methods Of Feature Selection:

A. Univariate Selection

This method uses the SelectKBest class along with a group of other tests that analyze statistics to choose a certain number of features. The chosen features are those that are closely related to the output variable. The statistical test used, is the Chi-squared test (chi^2) for positive features to select the 5 most important features. The estimated accuracy is **77.01%.** So the selected features are TSH, TT4, T3,T4U,FTI.



```
Score list: [3.22136430e+02 1.01652619e+01 2.19382626e+01 9.18188737e-01
 3.31717994e+00 4.16650339e+00 2.18751443e+01 3.03820742e+00
 6.72378837e-01 1.38685312e+01 1.95354665e+01 9.61472903e-01
 2.65195332e+00 3.54047225e+01 1.65000000e+01 6.88244863e+00
 1.93007801e+00 1.53110468e+04 4.93159549e+00 1.34385046e+03
 6.48547557e-01 3.90479164e+03 1.05061398e+00 5.78684859e+02
 1.05061398e+00 1.82044258e+03]
Feature list: Index(['Age', 'Sex', 'On Thyroxine', 'Query on Thyroxine',
       'On Antithyroid Medication', 'Sick', 'Pregnant', 'Thyroid Surgery',
       'I131 Treatment', 'Query Hypothyroid', 'Query Hyperthyroid', 'Lithium',
       'Goitre', 'Tumor', 'Hypopituitary', ' Psych', 'TSH Measured', 'TSH',
       'T3 Measured', 'T3', 'TT4 Measured', 'TT4', 'T4U Measured', 'T4U',
       'FTI Measured', 'FTI'],
      dtype='object')
```

Fig. 2. Individual attribute scores

B. Recursive Feature Elimination (RFE)

In this method, an external estimator assigns weights or importance values to features and it recursively selects features creating a smaller and smaller set of features. Initially the estimator is trained on the entire set of features and it uses the ‘feature_importances_’ attribute or the ‘coef_’ attribute to test the significance of each feature. The ‘fit’ method provides information regarding the feature importance. Then the least important features are discarded from the dataset and these steps are performed repeated recursively until a specific number of desired features are reached. The estimated accuracy is **77.5%** and the 5 best features chosen by RFE are: TSH, T4U, TT4, T3, and FTI.

C. Tree Based Feature Selection

In this method of feature selection we use random forest classification to choose the most significant features. The weights are assigned to features, either by the ‘feature_importances_’ attribute or the ‘coef_’ attribute. The ‘fit’ method provides information regarding the feature importance. The random forest approach randomly chooses features at each iteration, therefore the sequence of feature importance list can change. Feature ranking is obtained by calculating the feature importances for every feature respectively. The feature ranking obtained, ranks TSH, TT4, T3, T4U, FTI and Age as the most important features of the dataset. The estimated accuracy is **76.7%.**

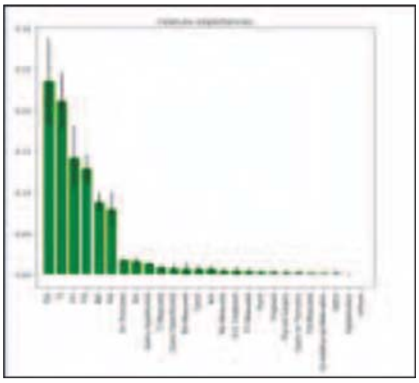


Fig. 3. Graph depicting importance of individual features

TABLE 1 . COMPARISON CHART OF FEATURE SELECTION TECHNIQUES

S no.	Name Of Technique	Accuracy of Feature Selection
1.	Univariate Selection	77.01%
2.	Recursive Feature Elimination	77.5%
3.	Tree Based Feature Selection	76.7%

So, the RFE technique gives the maximum accuracy.

D. FEATURE EXTRACTION

Feature extraction is the process of converting the original data into a dataset which has a minimal number of variables , containing only discriminatory information. It reduces the amount of input data by distilling its representative descriptive attributes.

Principal Component Analysis (PCA) for Feature Extraction:

Principal Component analysis or PCA is a procedure which utilizes a certain number of transformation procedures to transform a dataset of closely related variables into a set of variables that are uncorrelated known as principal components. The data is transformed in a way such that the first principal component has the greatest variance which implies that it accounts for the maximum amount of variability in the data. PCA is utilized as a tool for data analysis and for making models for prediction. It reveals the internal details of the data and gives an explanation for the variance in the data. The total variance is the sum of variances of all principal components. The fraction of variance explained by PCA is the ratio between that principal component and the total variance. So the variances of all principal components are divided by the total variance.

On applying PCA we get:

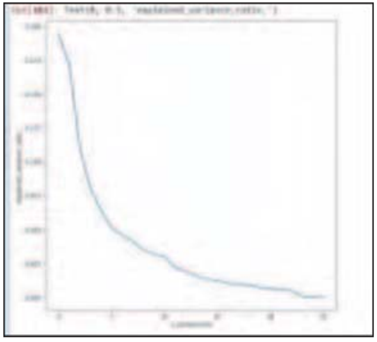


Fig. 4. Graph depicting total variance ratio computed by PCA

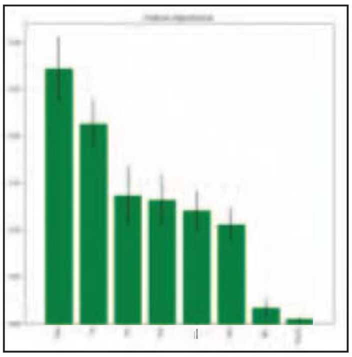


Fig. 5. Graph depicting importance of selected attributes

E. EXPERIMENTAL RESULTS USING WEKA

Waikato Environment for Knowledge Analysis also known as WEKA is a bunch of machine learning techniques for data mining problems. The techniques can be applied to a dataset or invoked from your own Java code. **WEKA possesses**

tools for classification, regression, dataset preprocessing clustering, and visualization.

TABLE 2. COMPARISON CHART OF FEATURE SELECTION TECHNIQUES OBTAINED FROM WEKA

S no.	Name Of Algorithm	Correctly Classified instances (percentage)	Incorrectly Classified instances (percentage)
1	Naïve Bayes	63.07%	36.92%
2	Multiclass Classifier	74.14%	25.85%
3	Random Forest	89.23%	10.26%

From these results it can be concluded that Random Forest is the best classification approach for this dataset. However this is a tentative analysis. The actual results were obtained after thoroughly preprocessing the dataset and then implementing classification techniques

V. CLASSIFICATION TECHNIQUES

A. Naïve Bayes Algorithm

It is a classification algorithm based on Baye's Theorem assuming independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naïve Bayes model is easy to create and helpful for very large data sets. Naïve Bayes is a simple technique yet it possesses the capability to outperform highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c) as given in the equation.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes). **P(c)** is the prior probability of class. **P(x|c)** is the likelihood which is the probability of predictor given class. **P(x)** is the prior probability of predictor.

Important Features Of The Naiive Bayes Algorithm

i. The features are independent of each other .Each feature is given some weight or importance. None of the attributes is irrelevant and is assumed to be contributing equally to the outcome.

ii. Assumptions made by Naïve Bayes are not correct in real world situations.

The estimated accuracy of the Naïve Bayes Algorithm was 16.8% before applying feature selection. However the accuracy significantly improved after applying feature selection and was estimated to be **74.37%.** The algorithm gives the highest accuracy(74.37%) when RFE Technique is applied.

B. Support Vector Machines

SVM (Support vector machine) is a classifier which works by separating classes through a hyper plane . The input to the algorithm is a set of labeled training data (supervised learning) and the output is a graph separating new instances of data into the classes through an optimal hyper plane. The hyper plane is basically a line separating a plane into 2 parts, each class lies on either side of the line. It can be utilized for both regression and classification problems, although it is mostly used for classification problems. Every data point is plotted in an n-dimensional space , 'n' is the total no. of features and the value of each feature is the value of that particular coordinate on the graph Support vectors are essentially the coordinators of individual observations and a Support Vector Machine is the model that best separates the 2 classes of support vectors.

Example of Linear Hyper plane

Input-

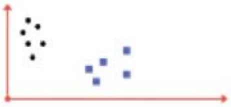


Fig. 6(a) The black dots and blue squares represent 2 different classes.

Output-

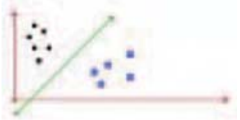


Fig. 6 (b) The green arrow in the output represents the optimal hyper plane separating the 2 classes.

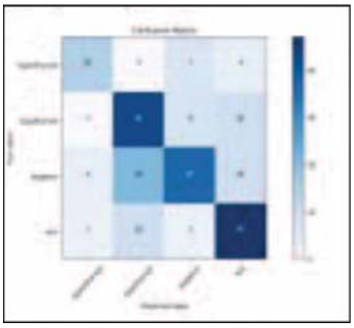


Fig. 7. Confusion matrix obtained by the SVM technique

i. **Advantages of SVM**

- It performs well with a clear margin of separation.
- Converts low dimensional spaces to high dimensional spaces.
- It uses memory efficiently.

ii. **Disadvantages of SVM**

- Requires higher training time hence does not work on larger datasets.
- Doesn't perform well with overlapping target classes.

The accuracy of the SVM technique improved significantly from 63.9% to 92.92% after applying feature selection.
The SVM technique gives the highest accuracy (92.92%) with the RFE method of Feature Selection.

C. Random Forest Algorithm

The Random Forest Approach is a classification as well as regression technique which can be used like bootstrapping with a number of decision trees making a forest. It is an optimal mix of tree predictors where every tree relies on the randomly selected values of the given vector. Once some new input data is received, the algorithm makes a decision tree for that data and places it in the forest containing the rest of the decision trees. The Random Forest technique provides improvement from the classical single decision tree approaches like CART and C4.5.

Advantages of Random Forest Approach

i. It has accuracy as good as Adaboost.
ii. Its faster than bagging or boosting.
iii. It is robust to noise.
iv. It is simple and easily parallelized.
v. It gives helpful error estimates

Steps to carry out this Algorithm

i. Select the total number of trees (Tn) to grow.
ii. Select the total no. of input variables (Vm) to split the nodes of the decision tree.
iii. Make the trees (decisions) grow.

For every tree the following is done:

- A sample of size 'S' is made from the 'n' no. of training cases and are made to grow.
- While making a decision tree grow at all the nodes, 'Vm' variables are chosen at random from 'M' (which is used to find the best split).
- The tree is grown to a maximum extent where there is no pruning.

The classification point 'K' is responsible for the collection of votes from every tree in the forest and then the majority of votes is used to decide on the class label

Mathematical Modelling

For every decision tree, the Sci-kit learn library calculates a node importance using the Gini importance , that assumes a binary tree(only 2 child nodes):

$$ni_j = w_j c_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)}$$

ni sub(j)= the importance of node j, **w sub(j)** = weighted number of samples reaching node j ,**C sub(j)=** the impurity value of node j, **left(j)** = child node from left split on node j, **right(j)** = child node from right split on node j
The importance of features on decision tree is calculated by:

$$fi_i = \frac{\sum j;node\ j\ splits\ on\ feature\ i\ ni_j}{\sum k \in all\ nodes\ ni_k}$$

fi (i) = the importance of feature i, **ni (j) =** the importance of node j.
These values are to be normalized to a value between 0 and 1 by dividing it by the sum of all feature importance values as follows:

$$normfi_i = \frac{fi_i}{\sum j \in all\ features\ fi_j}$$

The final feature importance at the random forest level is its average over all the trees. It can be calculated by dividing the sum of feature importances by the total number of trees.

$$RFfi(i) = \frac{\sum j \in all\ treesnorm_{fi(ij)}}{T}$$

RFfi (i) = the importance of feature i calculated from all trees,
normfi (ij)= the normalized feature importance for i in tree j,
T = total number of tree
The estimated accuracy of the algorithm is 78.21% (less than expected by the WEKA tool). The accuracy has improved from 75.5% to 78.21% after applying feature selection.
It gives the highest accuracy with the RFE technique of feature selection.(78.21%)

VI. PERFORMANCE ANALYSIS

TABLE 3. COMPARISON OF CLASSIFICATION TEACHNIQUES WITH THE RFE TECHNIQUE OF FEATURE SELECTION

S no.	Name Of Technique of Feature Selection	Accuracy of Feature Selection	Name Of Classification Technique	Accuracy on applying Classification algorithm
1.	RFE	77.5%	Naïve Bayes	74.37%
2.	RFE	77.5%	Random Forest	78.21%
3.	RFE	77.5%	SVM	92.92%

Hence the SVM technique along with RFE technique of Feature Selection gives the highest accuracy of 92.92%.

VII. CONCLUSION

The combination of the Recursive Feature Elimination and the Support Vector Machine Technique has proven to be effective on this dataset. The feature set finally used is ‘Age’ , ‘Sex’ , ‘TSH’ , ‘TT4’ , ‘T4U’ , ‘T3’ , ‘FTI’. **Age** and **Sex** have been considered as important features because Thyroid disorders are said to occur during a particular age range (17-54 years) and is more prevalent in females. **TSH** (Thyroid Simulating Hormone) , it is a hormone that simulates the Thyroid gland to produce thyroxine and it is tested to check whether the thyroid gland is overactive or underactive. The normal range of TSH is between 0.4- 4.0(mU/L) . **TT4** is test to check thyroxine or T4 levels in the blood , the normal range of which lies between 5.0 – 12.0(ug/dL). **T4U** is a test to check the T4 levels that are bound to proteins and that prevent it from entering the various tissues that require thyroxine. **FTI** or Free T4 is a test to measure free thyroxine available in the blood that enters the various target tissues and exerts its effects. Hypothyroid patients have excess of FTI and Hyperthyroid patients have low level of FTI. **T3** (Triiodothyronine) is a test to measure the T3 levels in the body. Most of the T3 in the blood is produced by various body parts. A free or total T3 test is used to assess the function of the Thyroid gland.

This Support Vector Machine classification technique along with the Recursive Feature Elimination method of feature selection giving an accuracy of 92.92% hence proposed can help medical experts in decision-making and classifying thyroid patients into 4 classes of Thyroid disorders namely Hypothyroid, Hyperthyroid, Sick Euthyroid and Euthyroid (negative). In the diagnosis of Thyroid diseases, the accurate interpretation of data along with expert clinical examination and investigation is a significant issue.

REFERENCES

[1] Prerana, P.S. and Taneja, “Predictive data mining for diagnosis of thyroid disease using neural network”, Int J Res Manage Sci Technol, 3(2), pp.75-80 , April 2015, in press.

[2] Chen, L., Li, X., Sheng, Q.Z., Peng, W.C., Bennett, J., Hu, H.Y. and Huang N. , “Mining health examination records—A graph-based approach”, IEEE Transactions on Knowledge and Data Engineering, 28(9), pp.2423-2437, 2016, in press.

[3] Ammulu Venugopal , “Thyroid data prediction using data classification algorithm”. International Journal, 4, pp.208-212, 2017 , in press.

[4] Shankar, K., Lakshmanaprabu, S.K., Gupta, D., Maseleno, A. and de Albuquerque, V.H.C,“ Optimal feature-based multi-kernel SVM approach for thyroid disease classification”. The Journal of Supercomputing, pp.1-16. , 2018 , in press.

[5] Ioniţă, I. and Ioniţă, L,“ Prediction of thyroid disease using data mining techniques”, Broad Research in Artificial Intelligence and Neuroscience, 7(3), pp.115-124 , 2016 , in press.

[6] Kulkarni, S.N. and Karwankar, A.R , “ Thyroid disease detection using modified fuzzy hyperline segment clustering neural network” , International Journal of Computers & Technology, 3(3b), pp.466-469 , 2012 , in press.

[7] Pal, V.K , “ An intelligent system for diagnosing thyroid disease in pregnant ladies through artificial neural network”,2019, in press.

[8] Ahmed, J. and Soomrani, M.A.R, “TDTD: Thyroid disease type diagnostics” In 2016 International Conference on Intelligent Systems Engineering (ICISE) (pp. 44-50). IEEE , 2016 , in press.

[9] Pandey, S., Tiwari, A., Shrivas, A. K., & Sharma, V, “Thyroid classification using ensemble model with feature selection”, Int J Comput Sci Inf Technol, 6(3), 2395-2398, 2016 , in press.

[10] Temurtas, F, “A comparative study on thyroid disease diagnosis using neural networks” Expert Systems with Applications, 36(1), pp.944-949, 2019, in press.