Comparative Analysis of Thyroid Disease based on Hormone Level using Data Mining Techniques

Pushpanathan G [1],
Asst. Professor,
Dept. of CSE, Cambridge
Institute of Technology, Bengaluru.

Gowthami Singh [2], U Anil Kumar [3],
Puneetha Ramesh [4], Anuj Kumar Dubey [5]
Student, Dept. of CSE,
CiTech, Bengaluru.

***Abstract*— Classification is one of the most important technique in machine learning. It has been found that classification is most widely used in all sectors. Classification is supervised learning technique which uses predefined data set to make accurate decisions. In this work, we use techniques such as SVM, KNN, Decision tree and Naïve bayes, Random forest and Logistic Regression to identify the type of thyroid disease using ANACONDA as software and python programming language is used to implement these algorithms. As last step, we compare accuracy of logistic regression and random forest and represent in graphical form. This system is used to provide diagnosis report of thyroid to patients with reduced cost.**

Keywords*— KNN, Naive bayes, ANACONDA, Decision tree, Random forest, Logistic Regression.

I. INTRODUCTION

Classification techniques plays an important role in analyzing diseases with reduced cost to the patients. Recently, Disease diagnosis is the difficult step in medical field because numerous diseases occurs every year. Now-a-days, a wide variety of diseases are detected worldwide. So, the detection of type of disease has become crucial. Today, Diseases are increasing exponentially and new variety of diseases are discovered. So, there is need of system to identify the type of disease and also to detect the type of disorder. Data mining plays a vital role in dealing with disease diagnosis. Thyroid is one of disease which has spread throughout the world. As per recent research, thyroid has wide spread all over the world and women are likely to be affected more to thyroid disease than men. It is caused due to dysfunction of thyroid gland located at Adam's apple of human body.

II. INFORMATION ABOUT THYROID

The thyroid is an endocrine gland which secretes the hormones directly into the blood. The capacity of thyroid organis to create the thyroid hormones. Thyroid disease most likely to affects the women during pregnancy and menstrual cycle. These thyroid hormones will control the metabolism of body. If these hormones are not produced in proper quantity, it slow down the overall body processes. Thyroid is mainly of two types - Hyperthyroidism and Hypothyroidism. These are caused due to production of thyroid hormones inappropriately. Thyroid disease can even lead to cancer which may lead to death. Hyperthyroidism is increase in functioning of thyroid gland and vice versa. This is caused due to lack of iodine in human body which leads to thyroid and severe conditions like goiter, cretinism, myxedema. Iodine is an important element in human body which are mainly responsible to produce T4 and T3 hormones. The thyroid gland produces tri-iodothyronine (T3) and Lthyroxine (T4) which plays an important role in metabolism of the body and it is a material which is binded with protein. Thyroid hormones are required for mental and physical development of body. It is responsible for maintaining electrolyte and other mineral levels in body. It also controls central nervous system, brain and other parts of the body. So, Thyroid hormones is important for all over development of human body. Thyroid disorders are the condition which makes improper functioning of the thyroid gland. This gland plays numerous and diverse roles such as regulation of various metabolic processes in human body. The hypothalamus in the brain produces Thryrotropin releasing hormone (TRH) which causes pituitary gland to release hormone called thyroid stimulating hormone (TSH). This happens when there is a lack of production of thyroid hormone. TSH is the one which stimulates thyroid gland to release T4. Improper function of thyroid gland also affects the thyroid as well as other functions of the human body. Thyroid disease can be classified mainly into two types are –

1) *Hyperthyroidism*

This is a disorder where thyroid gland produces enormous amount of thyroid hormones. Common symptoms of this includes restlessness, agitation, tremors, weight loss, rapid heartbeat, frequent bowel movements. In this disease most of T4 hormone gets converted to T3. The main causes of hyperthyroidism is excessive intake of iodine, abnormal secretion of TSH, Excessive intake of thyroid hormones. Grave's disease is severe condition of thyroid which may lead to death.

2) Hypothyroidism

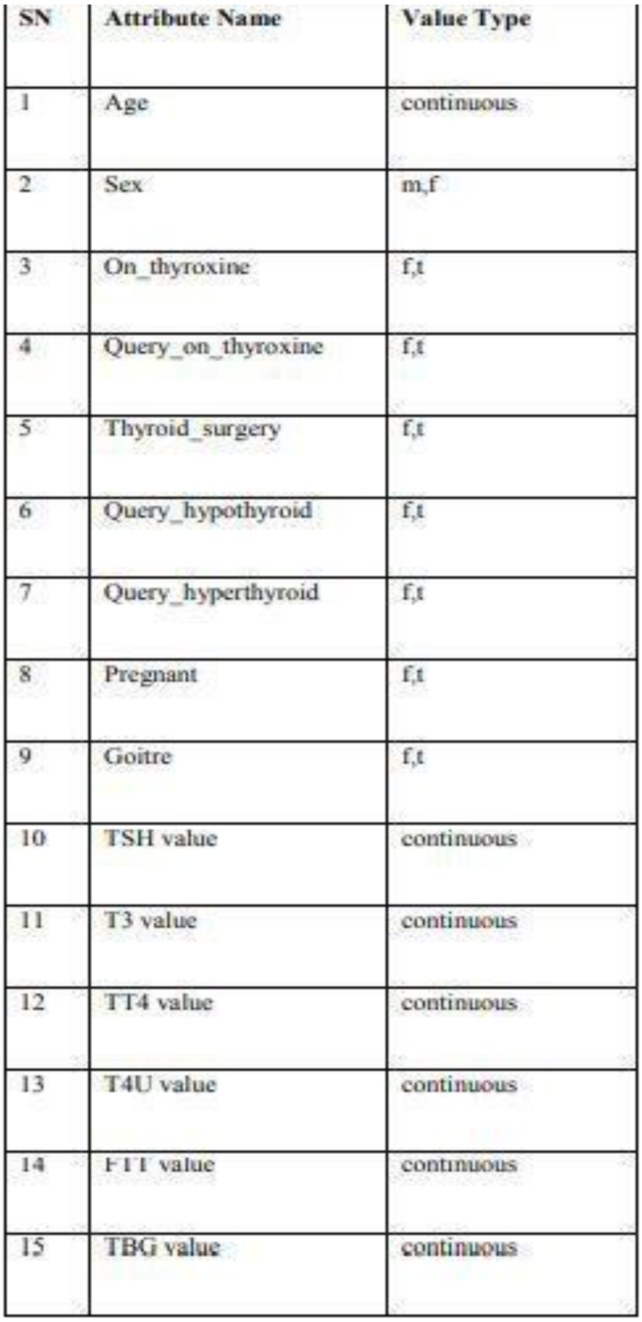
Hypothyroidism is a situation where amount of production of thyroid hormone reduces. Common symptoms of hypothyroidism includes dry skin, constipation, feeling cold, prolonged menstrual bleeding, sudden weight gain. This also causes other disorders such as Thyroid hormone resistance, Hashimoto's thyroiditis.

III. LITERATURE SURVEY

MC-CNN architecture can be used to detect thyroid nodules from the ultra sound images of thyroid disease. It can detect the thyroid nodules in smarter way. This system also plays an important role in detecting thyroid nodules in easier way. [5]. As compared to this reference we found a work which contained A research study which was made on different algorithms to check accuracy in detecting various types of cancer. This research included all variety of cancer such as colon cancer, thyroid cancer and compared the accuracy using area under the curve method. This study compared various boost algorithms with support Vector Machine algorithm. SVM is proved to be more efficient than xg boost, d boost and boost 1 algorithms. So, this study concluded that SVM is the best algorithm to detect the cancer [6]. Further another system was proposed with KNN and SVM for thyroid disease diagnosis. This system had two phases. In first phase, it is checked whether dataset contains any missing value if it has missing value then KNN imputation method is used to fill the missing values in input. And after that the dataset is sent into SVM to detect thyroid disease. In second phase, if dataset contains no missing value it is sent directly into SVM without sending data into KNN algorithm. This system is one which can reduce thyroid diagnosis cost for patients [2]. In year 2019, detection of thyroid nodule is done with Convolutional neural networks which takes ultrasound images as input [3]. Later a system was innovated which uses data mining techniques such as KNN, SVM, Decision tree, Back propagation techniques to predict thyroid disease [1]. A work on diagnosis of thyroid using Naive bayes classifiers, KNN, Decision tree and SVM and compared accuracy of all algorithms but decision tree performed highest accuracy of 98.89% than other algorithms [4].

IV. DATASET DESCRIPTION

Data set is collected from repository data which consists of records of thyroid patients. Thyroid data set consists of both boolean or contiunous valued variables. This dataset has fifteen attributes



SN	Attribute Name	Value Type
1	Age	continuous
2	Sex	m,f
3	On_thyroxine	f,t
4	Query_on_thyroxine	f,t
5	Thyroid_surgery	f,t
6	Query_hypothyroid	f,t
7	Query_hyperthyroid	f,t
8	Pregnant	f,t
9	Goitre	f,t
10	TSH value	continuous
11	T3 value	continuous
12	TT4 value	continuous
13	T4U value	continuous
14	FTI value	continuous
15	TBG value	continuous

Table 1. Dataset description

V. DATA MINING TECHNIQUES USED IN THE STSTEM

1) K – Nearest Neighbors

KNN is unsupervised learning algorithm which is simple algorithm that compares test data with existing data set to find missing attribute values in test data. It is mainly used in applications such as economic forecasting, data compression, genetics. The main advantage is easy to implement, robust to noise in input data. On other hand it takes more time to classify the data as it processes the entire training data.

2) Support Vector Machine

It is a supervised learning algorithm which is used for both classification as well as regression. It uses hyperplane which separates the training data into separate classes. Advantage of SVM is it uses less memory because it uses subset of training data. But it has disadvantage of using more dimensional space if the number of features increase. Kernel function can be used to overcome this drawback.

3) Decision Tree

Decision tree is a supervised tree algorithm which help in decision making by using tree presentation. Each leaf node is class label, internal node is attribute and root node represents resultant output. Decision tree splits nodes on all available variables and selects best split to get final outcome.

4) Random Forest

Random Forest is a flexible algorithm which consists of many decision trees. This algorithm constructs decision trees for data samples and then takes predictions from all decision trees and then combines all decision trees to select best among the decision trees. This works in similar way to voting and finally selects the best decision based on majority.

5) Logistic Regression

This is appropriate algorithm for Regression analysis which explains relationship between one dependent binary variable and one or more ratio level independent variables. This algorithm best suited for continuous variables. Advantage of Logistic Regression is easier to implement and efficient to train.

VI. PROPOSED WORK

The proposed system has used Random Forest and Linear/Logistic Regression Techniques to classify the thyroid dataset. The thyroid Dataset is taken from a site which has thyroid patient records. The current system has three phases in classifying thyroid data sets. This includes Input phase, pre-processing and Classification phase. The input phase is where the predefined data set collected from repository. Next phase is pre-processing phase where missing values are filled and last phase is where Random forest as well as logistic regression are used to classify thyroid data into its types. The output is three types like hypothyroid, normal and hyperthyroid as shown in fig 3. At last the performance is calculated and compared to identify the best classification algorithm as shown in fig 4.

Advantages of the system

We will show more accuracy with the proposed work in terms of performance.

We applied preprocessed technique to minimize the execution time.

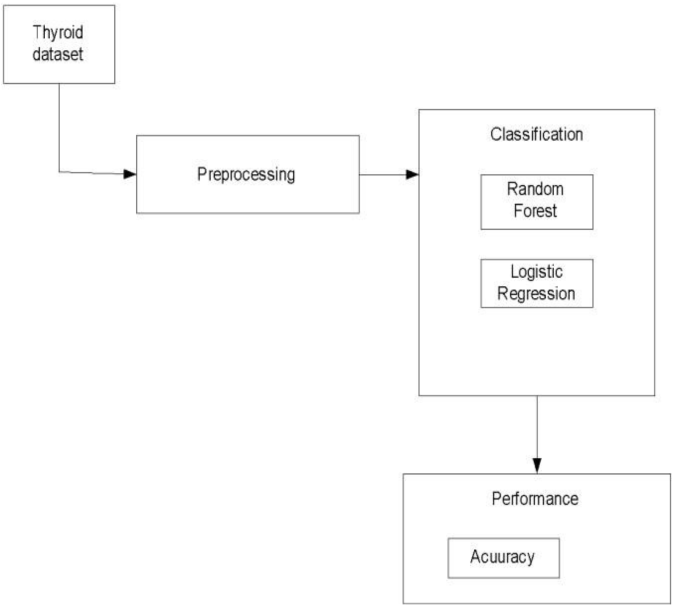


Fig 1. System for thyroid classification

VII. IMPLEMENTATION

a) Random Forest

1. Select k data sets randomly from given training data
2. For selected data sets build a decision trees for these subsets which is chosen.
3. Choose number N for decision trees that you want to build.
4. Repeat steps 1 and 2.
5. For new data sets, calculate each decision tree prediction and add new data set to class which has majority.

b) *Logistic Regression*

1. Draw the scatterplot. Look for 1) linear or non-linear pattern of the data and 2) deviations from the pattern If there are outliers, you may consider removing them only if there is a non-statistical reason to do so.

2. Fit the least-squares regression line to the data and check the assumptions of the model by looking at the Residual Plot and normal probability plot If the assumptions of the model appear not to be met, a transformation may be necessary.

3. If necessary, transform the data and re-fit the least-squares regression line using the transformed data.

4. If a transformation was done, go back to step 1. Otherwise, proceed to step 5.

5. Once a good-fitting model is determined, write the equation of the least-squares regression line. Include the

standard errors of the estimates, the estimate of σ , and R-squared.

6. Determine if the explanatory variable is a significant predictor of the response variable by performing a t-test or F-test. Include a confidence interval for the estimate of the regression coefficient (slope).

VIII. RESULTS

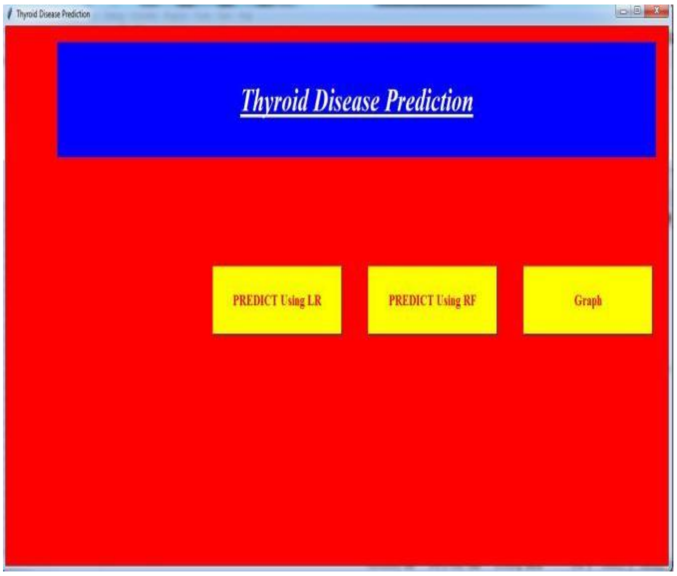
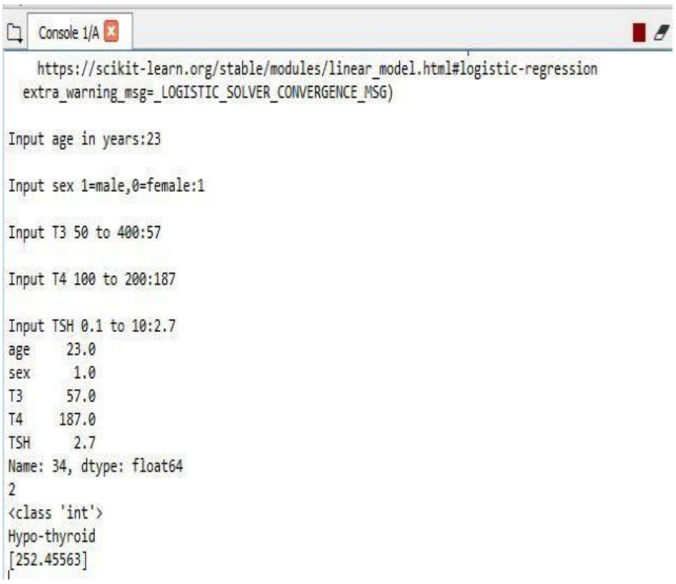


Fig 2. Figure of Main screen of the model

Above figure represents the Main Screen which has front end with various buttons to classify the output.



```
   https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG)

Input age in years:23

Input sex 1=male,0=female:1

Input T3 50 to 400:57

Input T4 100 to 200:187

Input TSH 0.1 to 10:2.7
age     23.0
sex      1.0
T3      57.0
T4     187.0
TSH      2.7
Name: 34, dtype: float64
2
<class 'int'>
Hypo-thyroid
[252.45563]
```

Fig 3. Figure of result screen of Regression model

Above figure has classified results for the given inputs using regression model. Above figure has classified results for the given inputs using random forest classifier model.

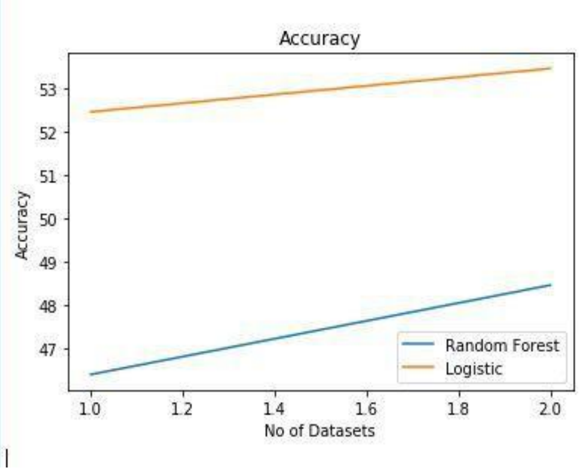


Fig 4. Figure of performance and comparison

Above figure has accuracy comparison graph for both regression and random forest classifier mode.

IX. CONCLUSION

In this work, we have used machine learning algorithms to predict thyroid disease. In this system, we have used data mining classification algorithms and regression algorithms. So, both regression and classification are combined to produce accurate diagnosis results. The logistic regression is more efficient and accurate compared to other classification techniques. But other recent techniques can be combined in future to give still more accurate results of thyroid diagnosis.

REFERENCES

[1] Bibi Amina Begum and Dr.Parkavi A "Prediction of thyroid disease using data mining techniques",5th International Conference on Advanced Computing & Communication Systems (ICACCS) 2019. *(references)*

[2] Aswathi A K and Anil Antony "An Intelligent System for thyroid disease classification and diagnosis" 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE Xplore Compliant - Part Number: CFP18BAC-ART; ISBN:978-1-5386-1974-2. *(references)*

[3] Shuaining Xie, "Thyroid nodule detection in ultrasound images with convolutional neural networks"14th IEEE Conference on Industrial Electronics and Applications (ICIEA) 2019. *(references)*

[4] Umar Sidiq , Dr. Syed Mutahar Aaqib , Dr. Rafi Ahmad Khan, "Diagnosis of various thyroid ailments using data mining classification techniques" ,International journal of Scientific research in Computer science ,Engineering and Information Technology 2019. *(references)*

[5] Wenfeng song, "Multi-task cascade convolution neural networks for automatic thyroid nodule detection and recognition "JOURNAL OF LATEX CLASS FILES, VOL.14,NO.8, AUGUST 2015.*(references)*

[6] Turki Turki, "An Impirical study of machine learning algorithms for cancer detection",978-1-5386-5053-0/18/$31.00 ©2018 IEEE. *(references)*

[7] K Saravana Kumar, Dr. R. Manicka Chezian,"Support Vector Machine and K- Nearest Neighbor based analysis for the prediction of hypothyroid", International Journal of Pharma and Bio Sciences" ,volume - 2,Issue - 5,page no-(447-453),2014.

[8] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3169866/(accessed dec 2015)., in press.

[9] G. Zhang, L.V. Berardi, An investigation of neural networks in thyroid function diagnosis, Health Care Manage. Sci. (1998).,in press.

[10] Xia C, Hsu W (2006) BORDER: efficient computation of boundary points. In: IEEE, 2006.
[11] http://en.wikipedia.org.Lastaccessed on Dec24.,in press.
[12] Apte& S.M. Weiss, Data Mining with DecisionTrees and Decision Rules,T.J.WatsonResearchCenter,http://www.research.ibm.com/dar/pap ers/pdf/fgcsaptewe issue_with_cover.pdf, (1997)., in press.
[13] Roychowdhury S (2014) DREAM: “diabetic retinopathy analysis using machine learning”. In:IEEE, 2014
[14] Chetty N, Vaisla KS, Patil N (2015) “An improved method for disease prediction using fuzzy approach”, In: IEEE, 2015
[15] S. SathiyaKeerthi, Olivier Chapelle, Dennis DeCoste "Building Support Vector Machines with Reduced Classifier Complexity" Journal of Machine Learning Research, Vol: 7, PP 1493- 515, January - (2006).
[16] Shen X, Lin Y (2004) “Gene expression data classification using SVM-KNN classifier". In: IEEE,2004.
[17] www.anaconda.com., in press.
[18] F. Temurtas, "A comparative study on thyroid disease diagnosis using neural networks," Expert Systems with Applications, vol. 36, 2009, pp. 944-949.
[19] G. H. John and P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers”,proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, San Francisco, 1995, pp.338-345.
[20] Ozyılmaz, L., Yıldırım, T. (2002). Diagnosis of thyroid disease using artificial neural network methods.In Proceedings of ICONIP’02 9th international conference on neural information processing (pp. 2033-2036). Singapore: Orchid Country Club.
[21] Polat, K., Sahan, S., &Gunes, S. (2007). A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted preprocessing for thyroid disease diagnosis. Expert Systems with Applications, 32, 1141-1147., in press.
[22] Sehgal MSB, Gondal I (2014) K-ranked covariance based missing values estimation for micro array data classification. In: IEEE, 2004.
[23] Bonner A (2004) Comparison of discrimination methods for peptide classification in tandem mass spectrometry.In: IEEE, 2004..
[24] HalifeKodaz et al. Medical application of information gain based artificial immune recognition system (AIRS): diagnosis of thyroid disease., in press.
[25] Joel Jacob et al. "Diagnosis of Liver Disease Using Machine Learning Techniques". (IRJET) Volume: 05 Issue: 04 | Apr-2018
[26] .K. Pavya et al. "Diagnosis of Thyroid Disease Using Data Mining Techniques: A Study". (IRJET)Volume: 03 Issue: 11 | Nov -2016.
[27] Keles, A., and Keles, A., ESTDD: Expert system for thyroid diseases diagnosis. Expert Syst. Appl.34(1):242-246, 2008., in press.
[28] Dogantekin, E., Dogantekin, A., and Avci, D.,An expert system based on generalized discriminant analysis and wavelet support vector machine for diagnosis of thyroid diseases. Expert Syst. Appl. 38(1):146-150, 2011., in press.
[29] M.P.Gopinath "Comparative Study on Classification Algorithm for Thyroid Data Set",.International Journal of Pure and Applied Mathematics Volume 117 No. 7 2017, 53-63.