Thyroid Prediction using Machine Learning Techniques

Anamitra Maji
MT19112
M.Tech CSE, IIIT Delhi

Pragya Dara
MT19126
M.Tech CSE, IIIT Delhi

Sameeksha Gupta
MT19096
M.Tech CSE, IIIT Delhi

Swati Verma
MT19073
M.Tech CSE, IIIT Delhi

***Abstract*—In the medical field, one of the challenging tasks is to diagnose a disease at the correct time so that doctors can provide proper treatment to the patients at the early stage of the disease. Thyroid is one of the very common diseases worldwide whose symptoms differ from person to person, and if left untreated, they get to worsen over time. More than half of the Indian population suffers from undiagnosed or misdiagnosed thyroid diseases. Almost a quarter of all men and half of all women in India die with evidence of an inflamed thyroid. So a prior diagnosis of the thyroid is very important and beneficial for the betterment of human life. Identifying the thyroid disorder from the laboratory test report is very complex and requires extensive knowledge and experience. Using machine learning techniques for this purpose eases the task of testing and makes it faster.**

Previously, various works have been done in predicting thyroid using different machine learning techniques, but for any normal user, it is not easy to understand the results of machine learning algorithms without having any prior knowledge about the disease. Also, as most of the users are from the medical field, they are mostly concerned about the results rather than understanding the algorithm behind predicting the disease. This paper proposes a solution by not only providing predictions with high accuracy but also by providing a user interface where any person can get knowledge about the disease and can get the predictive results by just filling out the form with some of the patient's information. Best five features are selected using K-fold cross-validation and different machine learning approaches like SVM, Decision Tree, Multilayer Perceptron Model, K nearest neighbors, Naive Bayes, and deep learning's Sequential Model are applied. All the models are compared on the basis of accuracy. It is found that the Decision Tree with height three and Deep learning's Sequential Model outperformed with an accuracy of 97.23% and 97.14%, respectively.

A website has also been developed which contains information about thyroid, a help page which will help the user to fill the form, a form where users can enter the major information about the patients, and a result page where the final result is obtained. For designing the front end, HTML, CSS, and JavaScript have been used. For connecting the front-end and back-end, Flask is used. Pickle library was used to save the model, which was then loaded in the python script containing the Flask code. The experimental results are shown with the help of this website.

So, our project provides not only very high accuracy but also gives important information about the disease with the help of a website, which makes it easier for the patients as well as the doctors to use our predictive model to their best use.

I. Introduction

While one out of ten adults in India experiences hypothyroidism, a current review led by Indian Thyroid Society portrays mindfulness for the illness positioned ninth when contrasted with other regular ailments [10]. Thyroid is a chronic disease that creates disorder in the immune system and might give rise to many other diseases.

Thyroid is a butterfly-shaped gland, which is located at the bottom of the throat responsible for producing two active thyroid hormones, levothyroxine (abbreviated T4) and triiodothyronine (abbreviated T3) into the bloodstream as the principal hormones. The main function of these hormones is to accelerate the human body metabolism, burn calories, protein, and restrict other hormonal glands when there is excessive secretion. These T3 and T4 are controlled by thyroid stimulating hormone (TSH) which is released by the brain. The pituitary gland acts like a thermostat to control the production of these hormones. When T3 and T4 are less in the bloodstream, then the pituitary gland releases more TSH and they are more, it controls the releasing of TSH. Hypothyroidism and hyperthyroidism are a result of an imbalance of thyroid hormones. Deficiency of thyroid hormones is referred to as **Hypothyroidism** and too much production of thyroid hormones is called **Hyperthyroidism**. Weight gain or failure to lose weight despite a proper weight loss regime, lethargy, reduced heart rate, increased cold sensitivity, numbness in hands, enlargement in the neck, dry skin and hair, could indicate hypothyroidism. Both types of conditions affect metabolism of the body. Hypothyroidism, in particular, causes a reduction in stroke volume and heart rate which further leads to lowered cardiac output along with a decrease in heart sounds. It also weakens the immunity of the body. 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. Another category in which thyroid diseases can be categorised into is **Sick Euthyroid**. Euthyroid is a condition which indicates proper functioning of thyroid gland. The level of hormones made by thyroid gland, i.e., T3 and T4, are comparatively higher in the patients suffering from sick euthyroid.

The large amount of data gathered from health care organizations has no organizational value unless transformed into most useful information and knowledge [11]. Data mining techniques come to our rescue when we want to make decisions regarding diagnosis of disease and give timely treatment to patients. Many works have been done in the field

	T3 level in blood	T4 level in blood	TSH level
Hyperthyroidism	normal	Too high	high
Hypothyroidism	normal	Very less	low
Sick Euthyroid	high	high	normal

Fig. 1. Classification of thyroid diseases

of diagnosis of thyroid diseases.

Some of the existing methods use the LDA (Latent Dirichlet Allocation) algorithm. LDA is a generative statistical model. It allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. But LDA doesn't work well if our data is not balanced and is very sensitive to overfitting. Serpen et al., used Learning Vector Quantizer (LVQ) and Probabilistic Potential Function Neural Network (PPFNN) for the test data [2]. Other works in this field include Multi-Layer Perceptron with back-propagation, adaptive conic section function neural networks(CSFNN) and Radial Basis Function(RBF) [3]. Multilayer Perceptron with LevenbergMarquardt algorithm was in [4]. Another work [5], have used a multi-kernel support vector machine model along with optimal feature selection for classifying thyroid patients. Use of MFHLSCNN (Modified Fuzzy Hyper Line Segment Clustering Neural Network) algorithm to classify thyroid patients is done in [6]. One of the significant works includes a comparison between various machine learning models for classification of thyroid disorders. [7] is a study on different features that are used in thyroid disease prediction using machine learning techniques.

Selection of the correct kernel function is a bit problematic because the performance of SVM depends greatly on the kernel function and the dataset used. Moreover, training a model of SVM involves higher computational cost and time. SVM gives the best result if our task is binary classification; for multi-class classification SVM is not suitable.

Use of Neural Networks gives best results if we have a lot of training data. Training any neural network model over a small set of training examples usually leads to overfitting of the model, despite the most optimal architecture being used.

Thyroid diseases are among the most well-known endocrine diseases in India. However, information on the predominance of thyroid diseases in India is generally inadequate [10].

We have used only five features to make as accurate a decision as possible. Although there are many researchers world-wide who have shown that neural networks or some other machine learning techniques outperform in the task of predicting hypothyroidism. But using a large number of features for this purpose is not that beneficial for doctors as well as patients. More features used, simply implies that the patient has to undergo several other clinical tests to get the values of those attributes in order to make a prediction of thyroid disease, which is in turn time consuming and not much cost effective. Our approach worked on both, reducing the number of features used to make a prediction along with building a model that gives accurate results.

II. Material and Method used

A. Data Description

The dataset used for prediction of thyroid is taken from UCI Machine Learning Repository containing 3772 records in training set and 3428 records in test dataset. The dataset has 21 features and one target or class label which contains three values: 1(normal), 2 (hyperthyroidism) and 3 (hypothyroidism). The link for the dataset is given below:

https://archive.ics.uci.edu/ml/machine-learning-databases/thyroid-disease/

A brief description of the features used while predicting thyroid is given below:

Age: Denotes the age of the patient
Sex: 0: Male, 1: Female
On Thyroxin: 0: False, 1: True
Query On Thyroxine: 0: False, 1: True
On Antithyroid Medication: 0: False, 1: True
Sick: 0: False, 1: True
Pregnant: 0: False, 1: True
Thyroid Surgery: 0: False, 1: True
I131Treatment:0: False, 1: True
Query Hypothyroid: 0: False, 1: True
Query Hyperthyroid: 0: False, 1: True
Lithium: 0: False, 1: True
Goitre: 0: False, 1: True
Tumor: 0: False, 1: True
Hypopituitary: 0: False, 1: True
Psych: 0: False, 1: True
TSH(Thyroid Stimulating Hormone):Its value ranges from 0.0 to 0.53.
T3(triiodothyronine):Its value ranges from 0.0005 to 0.18.
TT4(levothyroxine):Its value ranges from 0.0 to 0.53.
T4U:Its value ranges from 0.002 to 0.6.
FTI(Free Thyroxine Index):Its value ranges from 0.002 to 0.642.

B. Data Preprocessing

1) Data Cleaning: It is important to clean data before applying any data mining technique as presence of noise or missing values can give wrong predictions. So, in our project null values, missing values and presence of duplicate records were checked and it was found that there were no missing values or duplicate rows in the dataset taken for prediction. Categorical attributes were also checked whether they are properly encoded or not and it was found that categorical data are already encoded.

2) Class Distribution and Class Imbalance: Out of 3772 records present in the training set, 93 records belong to normal patients, 191 records belong to the patients having hyperthyroidism, and 3488 records belong to patients suffering from hypothyroidism. Similarly, out of 3428 records in the testing dataset, 73 records belong to normal patients, 177 records belong to the patients with hyperthyroidism, and 3178 records belong to patients with hypothyroidism. The class distribution for training and testing sets is shown below in figure 1 and 2. From the figure, it can be seen that there is no class imbalance problem because all classes are equally represented in the training and test set.

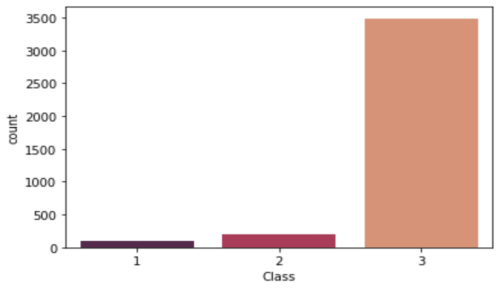

Fig. 2. Training Class Distribution

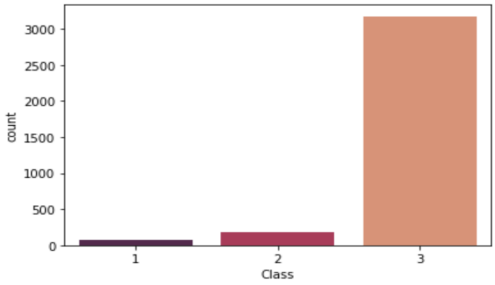

Fig. 3. Testing Class Distribution

C. Data visualization using PCA

For data visualization, Principal Component Analysis or PCA has been used, which selects two components from all the features, which gives maximum variation in the given dataset. It is very difficult to understand high dimensional data, so PCA is used to bring out useful patterns and highlight variations in a dataset. [1] From the graph shown in figure 4, it can be seen that the data is linearly separable.

D. Feature Selection

One of the essential concepts which significantly impacts the performance of any machine learning model is feature selection. Feature selection refers to selecting and removing irrelevant or partially relevant features which do not contribute much to the final output. Removal of such attributes not only reduces the training time but also increases the accuracy of the model. In this project, the correlation matrix is used for

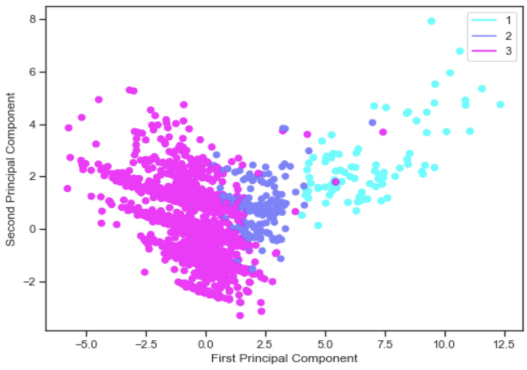

Fig. 4. Principle Component Analysis

extracting essential features. Correlation expresses how the features of a given dataset are associated with the target variable or each other. The correlation matrix for the dataset used is shown in the figure 5. It can be seen that only five features i.e, TSH(Thyroid Stimulating Hormone), T3(triiodothyronine), TT4(levothyroxine), T4U and FTI(Free Thyroxine Index) are correlated to each other. All the other features are not related either to each other or to the target label. So they are removed for further experiments.

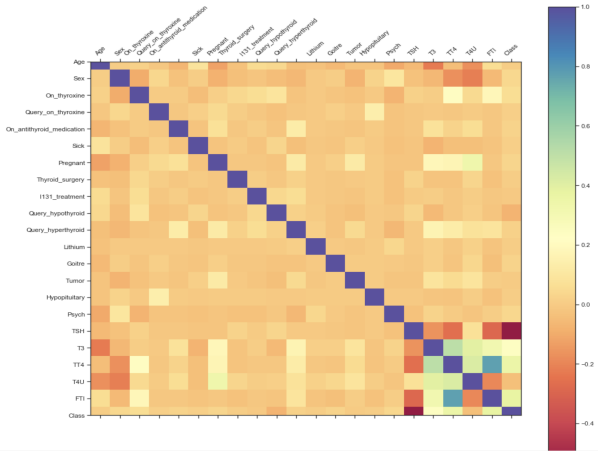

Fig. 5. Correlation Matrix

E. Data normalization using Standard Scaling

After selecting useful features, feature scaling is done to standardize the independent attributes existing in the dataset in a fixed range. Normalization is necessary to handle highly varying values in the data. For normalizing data, Standard Scalar has been used, which transforms the data in such a way that it has one standard deviation, and 0 mean. Standardization uses the formula:

$$X_{\text{new}} = \frac{X_{\text{i}} - X_{\text{mean}}}{StandardDeviation} \tag{1}$$

Where X_{new} is the scaled value of an attribute, X_{mean} is the mean of that particular column, and X_i is the current value of an attribute.

F. MACHINE LEARNING TECHNIQUES USED

1. SVM :

Support Vector Machine is a supervised machine learning algorithm that classifies data on the basis of a hyperplane in multidimensional space where each dimension represents a feature. Hyperplanes can be defined as decision boundaries that classify the data. The main objective of SVM is to select the optimal hyperplane, which has the maximum distance between data points of different classes. We used a 2-degree polynomial kernel for SVM.

2. Decision Tree:

One of the most used classifiers is Decision Tree, where the data is continuously split on the basis of certain features. It acts like a decision-making system represented in the form of a tree where each node represents a test condition, and edges represent outcomes that lead to the next non terminal node or leaf node. Leaf nodes represent the outcome or class label of the classifier. [13] The depth of the tree can be changed according to the requirement. Decision tree classifiers use impurity measures like entropy and information gain to take one decision at a time and follow the next path in the tree until the terminal node is reached.

3. Naive Bayes:

Naive Bayes is a probabilistic supervised machine learning classifier that is mostly used when the dataset is too large, like health-related datasets. The essence of the classifier is based on the Bayes Theorem, which uses conditional probability. This classifier works on a principle that every pair of attributes being classified does not depend on each other. [13] Given the attributes of a patient, the task of Naive Bayes Classifier is to classify whether a person is suffering from hyperthyroidism, hypothyroidism or the patient is normal.

4. K Nearest Neighbor

KNN is a supervised machine learning algorithm that can be used for classification as well as regression purposes. This algorithm works on the assumption that similar points exist in close proximity to each other. Choosing the correct value of k is a topic that is in much debate among researchers. So, to find 'k' that best suits our dataset, we tried different values of k and found out that k = 8 gives the best accuracy.

5. Multilayer Perceptron Model

MLPs are the classical feed-forward neural networks. It consists of one or more hidden layers, and predictions are made on the basis of the output layer, also known as the visible layer. In our implementation, we have used two hidden layers, each with 'Relu' activation function, a drop-out layer, and the output layer having 'softmax' as an activation function. [19]

G. Web Application

Our hypothyroid prediction model is also available as a web application. Making predictions using this application increases its usability and outreach among the individuals as well as organisations. The web application has one home page which gives brief information about the project,a help page which helps tell the user about important parameters that have been used for prediction and the range of input parameters. There is one form in which the user just has to input the values of T3, TT4, TSH, T4U, FTI and his/her prediction will be shown with the accuracy of about 98 percent using our model. We have used Flask library of python to develop a local website. The front end is developed using HTML, CSS and JavaScript. Flask is used to connect the front-end and the back-end. Pickle library was used to save the model which was then loaded in the python script containing the Flask code.

The home page of our website is shown below:

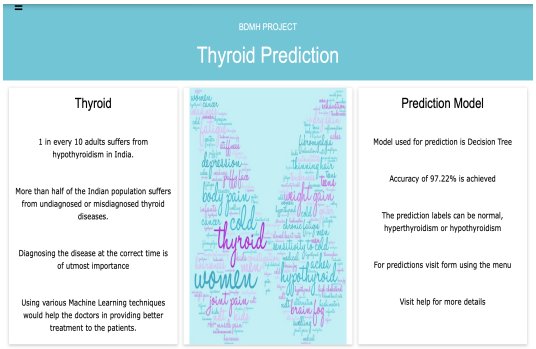

Fig. 6. Home Page of our website

III. EXPERIMENTS AND RESULTS

Among all the models, we found out that the decision tree performs the best in terms of accuracy measure for the given dataset. The training dataset used for building and evaluating our models using stratified k-fold cross-validation technique [17], where k =10. Here, splitting of data into folds is such that it ensures that each fold has the same proportion of observations with a given categorical value, such as the class outcome value.

Performance of different models can be summarised in the following table:

Training Model used	Training accuracy	Testing Accuracy	Normalisation
Sequential Model	97.9	97.14	yes
MLP	98.33	96.85	yes
Naive Bayes	95.25	94.78	yes
Decision Tree	98.22	97.23	yes
KNN	96.029	95.799	yes
SVM	93.75	92.7	no

Fig. 7. Machine Learning models and the respective highest accuracy

A graph has been plotted for comparing different models.

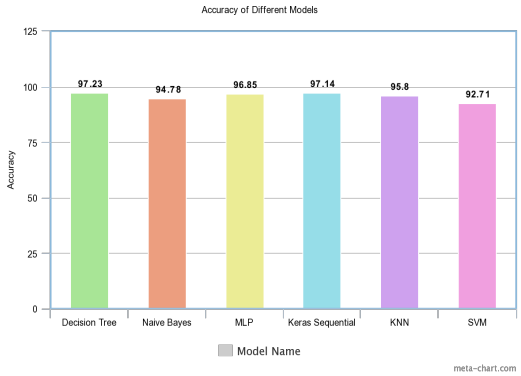

Fig. 8. Accuracy achieved by different models

It can be seen by the experiments that among all the models Decision tree and Sequential model performs best with the accuracy of 97.23% and 97.12% respectively. So, in our website also we are considering predictions from the decision tree only.

Graph between height and accuracy obtained at each height is plotted and we observed that at height = 3, training and testing accuracy is almost same. So, we can say that there is least overfitting and best accuracy at height 3.

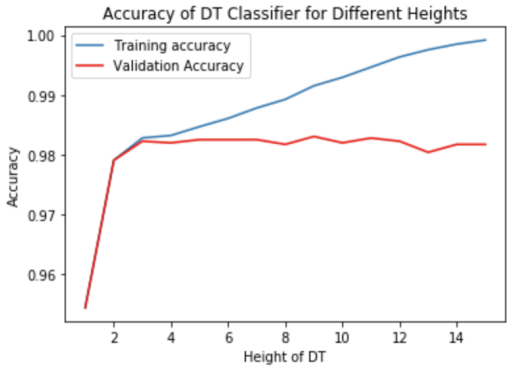

Fig. 9. Accuracy achieved for different heights in Decision Tree

Graph between accuracy and different values of k is plotted for KNN and it can be seen that the best results are obtained when k=10, because the training and testing accuracy have the least difference at k=10.

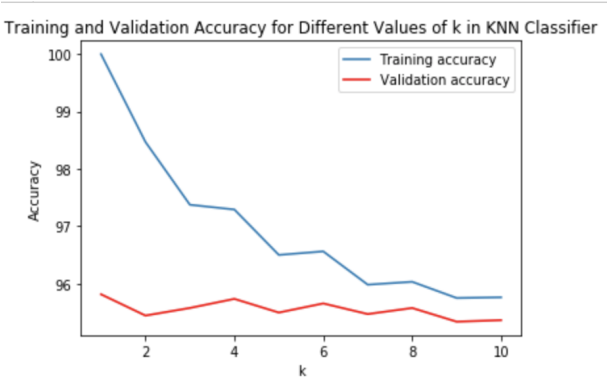

Fig. 10. Accuracy achieved for different values of k in KNN

Precision is the ratio

$$\frac{tp}{tp + fp} \tag{2}$$

where tp is the number of true positives and fp the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative.

Recall is the ratio

$$\frac{tp}{tp + fn} \tag{3}$$

where tp is the number of true positives and fn the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples.

F-beta score can be interpreted as a weighted harmonic mean of the precision and recall, where an F-beta score reaches its best value at 1 and worst score at 0. We have calculated this by taking beta = 1.

Precision, Recall and F-score was calculated for each model and the results are summarized below:

	Precision	Recall	F-1 Score	Testing Accuracy
SVM	0.9270	0.9271	0.92707	0.9271 (= 92.71%)
DT	0.9786	0.9734	0.9750	0.9734 (= 97.34%)
MLP	0.9765	0.9676	0.9702	0.9676 (= 96.76%)
KNN	0.9446	0.9536	0.9444	0.9536 (= 95.36%)
NB	0.9404	0.9486	0.9420	0.9486 (= 94.86%)

Fig. 11. Evaluation metrics for classification models

We can see that Decision Tree has the highest values of Precision, Recall and F-score.

A. Result form User Interface

A form is present in our website where the values of five main features T3, TT4, TSH, T4U and FTI are entered and in the next page prediction result is shown. Screenshots of the two web pages of website is shown below:

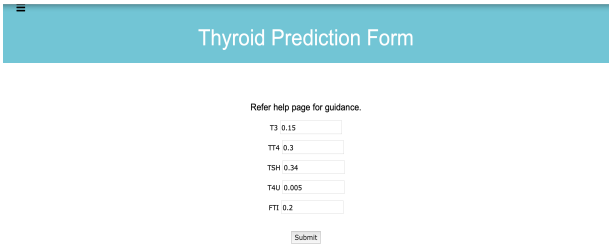

Fig. 12. Form taking values of T3, TT4, TSH, T4U and FTI

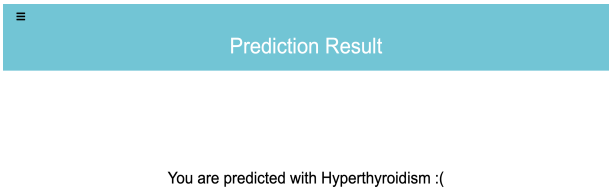

Fig. 13. Webpage giving the prediction

IV. Discussion and Future Work

In this paper different machine learning techniques along with the user interface for getting predictive result has been implemented. Using machine learning techniques for predicting thyroid eases the task of testing and makes it faster. It will further 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. The Decision Tree technique used along with the above-described feature selection and normalization techniques gives very much accurate predictions. This would definitely help medical practitioners in identifying thyroid patients and thus provide appropriate treatment.

In future we can extend our web application by adding some important feature like recommending doctors and medicines for different type of thyroids. We can also add some forms in which users can enter the symptoms and our system will be able to automatically recommend some exercise and home remedies which can help patients to maintain the level of hormones in their body.

V. Contribution of Each Author

- Anamitra Maji - Data cleaning and pre-processing, applying machine learning models
- Pragya Dara - Development of web application, applying neural network models
- Sameeksha Gupta - Feature Selection and applying machine learning models
- Swati Verma - Data Visualisation and applying machine learning models

References

[1] P. Duggal and S. Shukla, "Prediction Of Thyroid Disorders Using Advanced Machine Learning Techniques," 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2020, pp 670-675

[2] Ahmad, Waheed, et al. "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 (2017): 78-85

[3] Ozyilmaz, L. and T. Yildirim. Diagnosis of thyroid disease using artificial neural network methods. in Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on. 2002. IEEE.

[4] Temurtas, F., A comparative study on thyroid disease diagnosis using neural networks. Expert Systems with Applications, 2009. 36(1): p. 944-949.

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

[6] S.N. Kulkarni and A.R Karwankar, "Thyroid disease detection using modified fuzzy hyperline segment clustering neural network", International Journal of Computers & Technology, vol. 3, no. 3b, pp. 466-469, 2012.

[7] F Temurtas, "A comparative study on thyroid disease diagnosis using neural networks", Expert Systems with Applications, vol. 36, no. 1, pp. 944-949, 2019.

[8] Shahid, Afzal & Singh, Maheshwari & Raj, Rahul & Suman, Rashmi & Jawaid, Drakhshan & Alam, Muqtadir. (2019). A Study on Label TSH, T3, T4U, TT4, FTI in Hyperthyroidism and Hypothyroidism using Machine Learning Techniques. 930-933. 10.1109/ICCES45898.2019.9002284.

[9] Yadav, D.C., Pal, S. Thyroid prediction using ensemble data mining techniques. Int. j. inf. tecnol. (2019). https://doi.org/10.1007/s41870-019-00395-7

[10] Shankar, K., Lakshmanaprabu, S.K., Gupta, D. et al. Optimal feature-based multi-kernel SVM approach for thyroid disease classification. J Supercomput 76, 1128–1143 (2020). https://doi.org/10.1007/s11227-018-2469-4

[11] Saeed Shariati and Mahdi Motavalli Haghighi, Proposing Neural Network Efficient Training Model for Thyroid Disease Diagnosis, pp. 526-528, 2017.

[12] K. K. Mahurkar and D. P. Gaikwad, "Normalization using Improvised K-Means applied in diagnosing thyroid disease with ANN," 2017 International Conference on Trends in Electronics and Informatics (ICEI), Tirunelveli, 2017, pp. 579-583, doi: 10.1109/ICOEI.2017.8300768.

[13] A. Tyagi, R. Mehra and A. Saxena, "Interactive Thyroid Disease Prediction System Using Machine Learning Technique," 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan Himachal Pradesh, India, 2018, pp. 689-693

[14] Ahmad, Waheed & Huang, Lican & Ahmad, Ayaz & Shah, Farooq & Iqbal, Amjad & Saeed, Asma. (2017). Thyroid Diseases Forecasting Using a Hybrid Decision Support System Based on ANFIS, k-NN and Information Gain Method. Journal of Applied Environmental and Biological Sciences. 7. 78-85.

[15] S. Dash, M. N. Das and B. K. Mishra, "Implementation of an optimized classification model for prediction of hypothyroid disease risks," 2016 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, 2016, pp. 1-4, doi:10.1109/INVENTIVE.2016.7824794.

[16] Q. Pan, Y. Zhang, M. Zuo, L. Xiang and D. Chen, "Improved Ensemble Classification Method of Thyroid Disease Based on Random Forest," 2016 8th International Conference on Information Technology in Medicine and Education (ITME), Fuzhou, 2016, pp. 567-571, doi: 10.1109/ITME.2016.0134.

[17] Banu, Gulmohamed. (2016). Predicting Thyroid Disease using Linear Discriminant Analysis (LDA) Data Mining Technique. Communications on Applied Electronics. 4. 4-6. 10.5120/cae2016651990.

[18] Ionita, Irina. (2016). Prediction of Thyroid Disease Using Data Mining Techniques. BRAIN. Broad Research in Artificial Intelligence and Neuroscience. Vol.7. pp.115-124.

[19] Shivanee Pandey, Rohit Miri, S. R. Tandan, 2013, Diagnosis And Classification Of Hypothyroid Disease Using Data Mining Techniques, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH TECHNOLOGY (IJERT) Volume 02, Issue 06 (June 2013)

[20] Prerana, Sehgal, P., Taneja, K., Gharehchopogh, F.S. (2015). Predictive Data Mining for Diagnosis of Thyroid Disease using Neural Network.

[21] Ammulu and Venugopal . (2017). Thyroid Data Prediction Using Data Classification Algorithm. International Journal for Innovative Research in Science Technology, 4(2), 208-212.

[22] K. K. Mahurkar and D. P. Gaikwad, "Normalization using Improvised K-Means applied in diagnosing thyroid disease with ANN," 2017 International Conference on Trends in Electronics and Informatics (ICEI), Tirunelveli, 2017, pp. 579-583, doi: 10.1109/ICOEI.2017.8300768.