**Journal of Industrial Information Technology and Application** *ISSN(Online): 2586-0852*
Vol.10 No. 2

# Diagnosis of Thyroid Disorde Classification using Machine Learning Algorithm

*K.Balasree*[*]

Department of Computer Science SRM Arts and Science College, Kattankulathur

**Abstract**. Data mining is crucial in healthcare for early diagnosis and preventive treatment, particularly for long-term conditions. This study focuses on classifying hypothyroidism and hyperthyroidism using machine learning models to improve prediction accuracy.

Utilizing data from the UCI Machine Learning repository and the RapidMiner tool, the research evaluates four algorithms: Support Vector Machine (SVM), Random Forest, Naive Bayes, and Decision Trees. By analyzing classification and recall, the proposed model successfully identifies thyroid ailments with a 94% accuracy rate, demonstrating the effectiveness of these automated tools in clinical decision-making.

**Keywords;** Thyroid disease diagnosis, Decision Tree, Support Vector Machine, Naïve Bayes, Random Forest.

Cite this paper as: K.Balasree (2026) "Diagnosis of Thyroid Disorde Classification using Machine Learning Algorithm", Journal of Industrial Information Technology and Application, Vol. 10. No. 2, pp. 1337- 1350

[*] Corresponding author : Bala1996shree@gmail.com
Received: Jan. 7. 2026 Accepted: Feb. 19. 2026 Published: Jun. 30. 2026

# 1. Introduction

Thyroid disease is a very common disease that spreads widely in today's growing world. Because it is now so widely diffused, it often causes major health and life-threatening diseases in those who are infected. The functions that affected by thyroid gland is in results that turn the secretion of thyroid hormone. The symptoms that includes in thyroid are "Fatigue, Cold tolerance, Dry skin, Weight gain, Face swelling, menstrual cycles and Hair Fall" etc. Based on the statistics, the thyroid disorders are rising in India and discussing about the thyroid disease and also about the hypothyroid and hyperthyroid caused and symptoms are discussed. According to the survey, many peoples are affected by hypothyroid. Thyroid disease affects around 42 million people in developing countries like India, where it affects the majority of the population. Thyroid illness is discussed in this document, which includes information on hypothyroidism and hyperthyroidism. The machine learning algorithm is critical for forecasting the accuracy of the algorithm's outputs, which serves to lower the risk of thyroid disease in people.

### *1.1 Overview of Thyroid*

Thyroid looks like butterfly-shaped gland through the neck, which is situated at the above in the collarbone and one of the endocrine glands in charge of increasing hormone production Thyroid hormones are in charge of regulating the rates of a variety of body activities. The game has both fast heartbeats and calorie burning. In essence, these are the processes that occur inside the body's metabolism. Thyroid has two active thyroid hormones are presented; "levothyroxine (T4) and triiodothyronine (T3)".The body makes it when too much of hormone in thyroid, the thyroid gland which it makesmore hormones is called hypothyroid and little in thyroid hormone than the body need and it develops a condition which is called hyperthyroidism. The Thyroid problematic that includes the enlargement of goiter that includes in thyroid gland. The thyroid gland generates more thyroid hormones than the body needs, resulting in hyperthyroidism. Hypothyroidism doesn't makes enough thyroid compare to hyperthyroidism. It lumps the thyroid gland, Thyroid cancer and Thyroidnodules then the swelling of thyroid is thyroiditis.

There are many types and varieties of symptoms in thyroid, symptoms of the condition in thyroid is often very related to other conditions in medical and stages of life. This symptoms make sure that difficulties have related a thyroid issue. In thyroid symptoms, two groups are divided which is hypothyroid and hyperthyroid. Hypo have a little thyroid hormone and hyper have much than that the symptoms of an thyroid that includes: Gaining weight, Feeling Tired, Experiencing forgetfulness,

Frequent and heavy menstrual periods, Dry and course hair, Hoarse a voice and experiencing an intolerance to the cold temperature.

### *1.2 Hypothyroidism*

Hypothyroidism is underproduction of thyroid hormones.The energy production that requires the body's energy in amountof thyroid hormones and it drops the hormone that leads in theenergy as lower levels [15]. In hypothyroidism that has somecauses includes:

- Hashimoto's thyroiditis: It stops the hormones, whenthe body attacked by thyroid tissue finally then theydies. The thyroid has removed or surgically destroy,when removes in the thyroid gland.
- Iodine - Introduction to excessive: The dyes that givenbefore some X-rays might describe more iodine arethe medicine amiodarone, cold and sinus medicines.The greater risk for hypothyroidism is when you haveproblems in thyroid as past.
- Lithium: It is a drug that implicated in cause of hypothyroidism and it is for long term untreated time.The multiples of hormones system at early stage, which can lead the development. Most infants have thethyroid level checking routinely. For adults the hypothyroidism is due to these causes. Entirely most and lacks of gland in hypothyroid children unusually inactive and they are in quiet then it has a poor craziness and unreasonably it sleeps at long period times.

Thyroid gland nodules are very rare, accounting for just around 5% of all thyroid nodules. It has more than one thyroid nodules in several yearsthat to determine the cancerous. Radiation treatment in the neck and neck area early in childhood is more successful and has the potential to cure acne in those who are prone to acquiring thyroid cancer at a higher rate than the general population. The thyroid nodules are rarely required and itis about only 5% of nodules in thyroid been affected. The peoples received by the radiation treatment by head and neckearlier, remedies to acne and higher than normal risk of thyroidcancer.

### *1.3 Hyperthyroid*

Thyroiditis is a condition in which the thyroid gland is overactive, causing the thyroid to produce an excessive amount of hormones to circulate in blood is a condition is called hyperthyroidism. Hyper that means the over in Greek word. It includes the Thyroxine (T4) and Trithyroxine(T3). T4 is actually converted to T3 for bloodstream and thyroid hormones. T3 isthe hormone in thyroid that is in mostly active. The thyroid gland is a butterfly-shaped gland with a rounded appearance that is located at the front of the neck and the gland is regulated by pituitary gland itself in a brain and it regulates in pituitary gland as hypothalamus another itis in brain.

Symptoms of an overcharged thyroid hyperthyroidism:

- Feeling nervousness, anxiety and irritability.
- Consuming trouble sleeping.
- Losing the weight.
- Goiter or Enlarged thyroid gland.
- Weakness and tremors in muscles.
- Menstrual stop or irregular periods.
- Sensitive feel during heat.
- Problems having in vision or eye irritation.

Difference between hypothyroid and hyperthyroid:

Hypothyroidism is a condition that affected by people only a little bit in thyroid hormones or thyroid hormones is inunderproduction and hyperthyroidism is increasing population and affects much more in thyroid hormones.

## 2. Literature Review

Fatemeh Saiti et al[2] presents about the previous comparison with two classifiers: “Support vector Machines and Probabilistic Neural Network”. They have resulted as two parts both are similar but the last dataset that given better result and it includes 7000 attributes, SVM has been shown to perform much better in terms of accuracy.

Ebru Turanoglu-Bekar et al. paper suggested that to Learn how to apply the decision

tree technique for thyroid ailment classification in practise. To find the best classifier for thyroid sickness, currently existing decision tree algorithms were examined in terms of six performance criteria (ACC (percent), FME, MAE, PRE, REC, and kappa statistic).when compared to all algorithms Naïve Bayes gives the better accuracy which is 75%.

Lale OZYILMZ Tulay ULDIRIM., resulted about the Number of correct classification tested, Accuracy for classification and the Training epochs. These are the types of neural network used in this paper. CSFNN, it provides the classification's best rates and needs training times less than other algorithms. MLP gives the good rate in average but the same performance doesn't guaranteed single runs depends on random weight initialization in training and RBF is not the good choice for classifying the accuracy in such a problems[1].

Ankita Tyagi et al.shows that neural network about the overall techniques and analysis demonstrations that different technologies used on different papers showing different accuracies to endure the number of greater clinical tests with more effective cost and consuming about time[4].

Irina IoniŃă, Liviu IoniŃă Discussed about many algorithms such as "Naïve Bayes, Decision Tree, MLP and RBF Network" and Using the UCI repository, it is predicted that roughly 30% of Romanians are diagnosed with endemic goitre thyroid and the best classification model in all these we predicted is Decision Tree[3].

Bibi Amina Begum, Dr.Parkavi Predicted about the thyroid disease and to cut down on the quantity of noisy data in the patient's data from the database and algorithm is used to study the "Support Vector Machine, Decision tree and Naïve bayes", ID3 and finally it is used to predict and recover the thyroid patients and it helps to treat with better cost and effective for patients[7].

Yifei Chen et al., resulted about that there are suggested Xgboost classifier is used to two groups of diagnostics in order to acquire two groups of structures. Both the normal and sequence datasets provide better final results than the other. The accuracy is 92.11% and 92.54% correspondingly by applying the proposed method[8].

M. R. Nazari Kousarrizi et al., proposed about the classification and several methods that predicted in "support vector machine". It is used to predict the accurate value with two datasets. For the first dataset, the accuracy of classification is 98.62% by using SVM and subset is (2, 3,5). The highest accuracy reached in this for moreover the

structure has improved with 1.39% of accuracy with reported as best technique.

U. R. Acharya et al., results about the three different classifying techniques they are

KNN(K-Nearest Neighbour), Decision Tree, Probabilistic Neural Network(PNN). These are the classifiers compared with the performance in “Receiver Operating Characteristic (ROC)” curves and the results showed the performance measures with 0.987 under the ROC curve, then the accuracy classification is 98.9%, specificity is 99.8% and sensitivity 98%.

Randall S. Sexton et al., proposed about the Genetic algorithm and K-Nearest Neighbour algorithm and it has set the initial and momentum value, which is 1.0 and 0.9. This study does not prove the Genetic Algorithm that to dominate all problems but it designate Genetic Algorithm to enable manager’s use with Neural Network and more confidence with solution that reported. The Neural Network becomes a very powerful for all managers.

## 3. Methodology

### *3.1 Machine Learning Algorithms*

The machine learning algorithm is a powerful and efficient algorithm and it being with artificial neural network, it workswith structured and derive dataset at different sizes and a number of attributes larger dataset considering the vividnessof shortlisted algorithms in implementation to select the bestin classifying the thyroid dataset. In machine learning three types are commonly used. “Supervised Learning, Unsupervised Learning and Reinforcement Learning”.

#### *3.1.1 Decision Tree*

It is one of the supervised learning technique used in classification and regression problems and mostly solving theclassification problems preferably. In this decision tree, the two nodes are used, one is a leaf node another is decision node. Leaf nodes are output and the decisions not contains theadditional branches and Decision node which is used for multiple branches and make the decisions is located and presented in both leaf and decision node. Decision tree is a structured and it represents for dataset in future.

```
Algorithm: Decision Tree Training Process
        Training set = T;
        Attribute set: A
        Target Attribute = B;
        Split criterion =sB;
        Stop criterion =stop;
                Grow (T,A,C,sB,stop)
        If (stop (T) =false
        Then
                for all a_i ∈A
                        do find ai with the best sB(T)
                label current Node  with a;
                for all value x_i ∈ a
                        do label outgoing edge with x_i
                        T_sub =T where a = x_i;
                        crate subNode      = Grow (T,A,C,sB,stop)
        else current Node=left;
                label currentNode with c_i where c_i most common value
        of B ∈ T;
```

### *3.1.2 Support Vector Machine (SVM)*

SVM algorithms are a useful and efficient tool for classification and regression. Machine learning applications use SVM techniques as well. Its goal is to create decision boundary or to create a best line, the best in decision boundary is a hyper plane and the extreme points or vectors helps choosing to create is hyper plane. It plots the attributes in graph and broadly separates using boundary and nodes thatlies are name as support vectors. These are the two categorieswhich is hyper plane and boundary.

```
Algorithm: Support Vector Machine

Input: Thyroid Dataset S, including m patient
Pre-defined ϵ
        Train result set R , SV_i ∈R(i=1,…,m);
Output: multiple SVM Classified F_k
Execute:
        for each affected patent do
                for i=1,j=1 to m do
                        If D(SVi,SVj)< ϵ
                        SVj←SVi;
                        Update SVj;
                end if
        end for
```

### *3.1.3 Random Forest*

In machine learning and supervised learning methodologies, as well as other domains, it is one of the most extensively used algorithms. The big advantages in random forest used for both regression and classification problems. The name suggested "Random forest is a classifierthat has a number of decision trees on given dataset of various". Ensemble learning is the technique of combining many classifiers to tackle complex problems and enhance a model's performance. It is used to address difficult-to-solve issues and enhance the performance model. The eventual output of a random forest is predicted based on the majority of votes made for each prediction in the forest.

```
Algorithm: Random Forest
A training set S: = (x1,y1),……(xn,yn), Features F, and a
number of trees in forest B.
Function Random Forest (K,L)
        H ← ϕ
        For i ∈ 1, β do
                Ki ← A bootstrap sample from S
                hj ← Randomized Tree Learn (Ki,L)
                H  ← HU(hj)
        End for
        return H
end function
Function Randomized Tree Learn (K,L)
        At each node:
                F → Very small subset of L..
                        Split on best feature in f.
        Return the learned tree.
end function
```

### *3.1.4 Naïve Bayes*

This method, which is based on the Bayes theorem, is used to address classification problems that need supervised learning. The text classification is used mainly in high dimensional and one of the most simple and effective algorithm that is to buildfast machine learning. The probabilistic classifier. The classifier is probabilistic in this situation, and it has predicted the basis of probability, which is the object. Examples are Spam filtration, Sentimental analysis

Formula for Naïve Bayes: P(A|B = P(B|A)P(A)/P(B).

Algorithm: Naïve Bayes

$$P(A|B)= \frac{(B|A)P(A)}{P(B)}$$

$$P(B)=\sum P(B|A)P(A)$$

## 4. System Architecture

Figure 1 shows the below design in system to be proposed. The recommended system architecture is shown in the diagram below. The different components of the proposed system are as follows:

Data Pre-processing: The requirement that data be received and pre-processed into a usable format before being entered into a data order and verified for missing values during the pre-processing phase. The categorical data required transforming into numeric data thatis in case of missing values. The dataset applied in themachine learning algorithm, training set and tested in splitting.

Training Set: This data's are trained in the machine learning algorithm. The given data's are split to test using cost function in decision tree algorithm in the case of gini index. The numeric input variables are inserted in the thyroid datasetas support vector machine algorithm and it contains random model samples in trained data points to build trees and random subsets considers the node splitting.

Data Pruning**:** The clipping is concluded in the performance improvement called decision tree algorithm. The branches oftrees is difficulty in usage of removing and the data pruning increase the algorithm's accuracy and also it reduces overfitting. The complexity of function increases the SupportVector Machine speed.

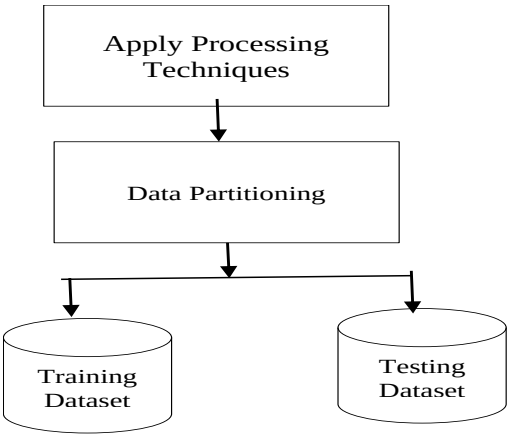


Figure 1. Training Data Pre Processing

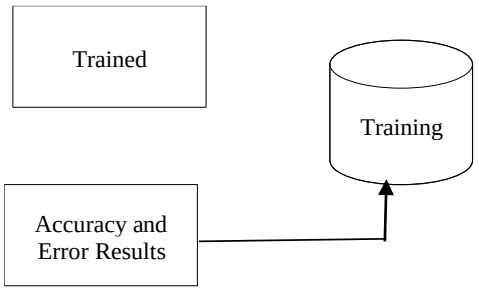


Figure 2. Data Pre Processing

## 5. Dataset Description

In order to perform the research, thyroid datasets were collected from the University of California, Irvine (UCI) machine learning repository and it contains 14 attributes and assigned with the various value

type according to the attribute represented and shown in table1 and in that table contains the Boolean values true/false. Therecords that included 6900 samples and 30 features which it contains a classes and has such a classifications mainly hypothyroidism, hyperthyroidism.

**Table 1. Dataset Description**

| Sl.No | Attribute Name | Value Type |
|---|---|---|
| 1 | Age | Continuous |
| 2 | Sex | Male, Female |
| 3 | On_Thyroxine | True, False |
| 4 | Query_On_Thyroxine | True, False |
| 5 | Sick | True, False |
| 6 | Pregnant | True, False |
| 7 | Thyroid Surgery | True, False |
| 8 | Query Hypothyroidism | Continuous |
| 9 | Query Hyperthyroidism | Continuous |
| 10 | TSH Value | Continuous |
| 11 | TT4 Value | Continuous |
| 12 | T4U Value | Continuous |
| 13 | T3 Value | Continuous |
| 14 | TBG Value | Continuous |

### *5.1 Evaluation Method:*

"F1 Score":

It is used to test and measure the accuracy. It is between precision and recall.

F1 = 2*1/1/precision*1/r

"For Precision":

Precision = TruePositives/FalsePositives

Recall = TruePositives/TruePositives/FalseNegatives

### *5.2 Result Analysis*

From all this above mentioned algorithms, the comparison observed in between all the algorithms and removed some datas after the result of that has predicted based on accuracy shown. The Support Vector Machine algorithm has 98.3% accuracy and it is the highest accuracy of all algorithms.and Naïve Bayes with 93.75%.

Figure 3 shows the Accuracy of statistics with classification models.

Table 2 represents the hypothyroidism accuracy statistics and table 3 represents

about the hyperthyroidism accuracy statistics with the use of Accuracy statistics with precision, recall and F measure

**Table 2. Hypothyroidism Accuracy**

| AccuracyStatistics | Naïve Baiyes | Decision Tree | SVM | RandomForest |
|---|---|---|---|---|
| Recall | 0.93 | 0.92 | 0.93 | 0.91 |
| Precision | 0.92 | 0.91 | 0.92 | 0.90 |
| F Measure | 093 | 0.91.2 | 0.92 | 0.91 |

**Table 3. Accuracy statistics with precision, recall and F measure**

| AccuracyStatistics | Naïve Baiyes | DecisionTree | SVM | RandomForest |
|---|---|---|---|---|
| Recall | 0.92 | 0.91 | 0.92 | 0.91 |
| Precision | 0.93 | 0.90 | 0.90 | 0.92 |
| F Measure | 0.94 | 0.92 | 0.91 | 0.90 |
| Accuracy | 93.7% | 91.2% | 92.3% | 90.6% |

## 6. Result and Discussion

Accuracy Statistics of Classification Models:

In accuracy statistics, the level of algorithm accuracy with using support vector machine, Decision tree, Random Forest and Naïve Bayes.

In this paper, the analyses about the impact of certain attributes over the accuracy in classification model. The following are the attributes which is removed are: Query_on_thyroxine, Query_on_Hypothyroid,Query_on_Hyperthyroid, it presents the accuracy. The classification modelbased on Naïve Bayes obtained with best accuracy as (93.7%), while Naïve Bayes obtained the weakest classification.

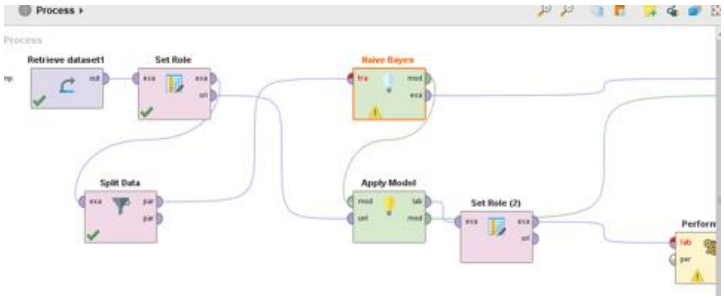


Figure 3. Naïve Bayes design

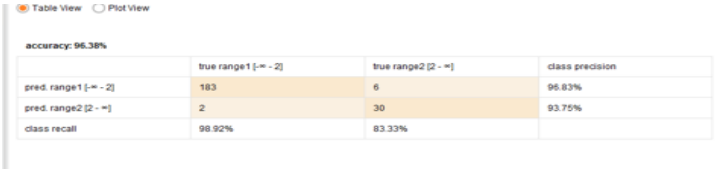


Figure 4. Naïve Bayes Accuracy Classification

Figure 3 and 4 represents the Naïve bayes result. Figure 4 shown above with the connecting procedure and 5 represents and shows about the results of Naïve Bayes with 93.7%.

## 7. Conclusion

In this paper the author discussed and applied four classification models (“Naïve Bayes, Support Vector Machine and Random Forest”). The dataset are taken from UCI Machine Learning repository, and it is used to predict the accuracy of precision and recall result using algorithm's. It shows results with rapid miner tool. In this, the best classification is Naïve Bayes with 93.7 accuracy and also classified with the effective data’s which helps to find the treatment with better cost and facilitates the management. The next study will concentrate on the elements that contribute to the identification of thyroid illness in a greater number of individuals and testing with more data mining techniques.

## References


[1] Ozyilmaz, L., & Yildirim, T. (2002, November). Diagnosis of thyroid disease using n artificial neural network methods. In Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP'02. (Vol. 4, pp. 2033-2036). IEEE.

[2] Saiti, F., Naini, A. A., Shoorehdeli, M. A., & Teshnehlab, M. (2009, June). Thyroid disease diagnosis based on genetic algorithms using PNN and SVM. In 2009 3rd International Conference on Bioinformatics and Biomedical Engineering (pp. 1-4). IEEE.

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

[4] Tyagi, A., Mehra, R., & Saxena, A. (2018, December). Interactive thyroid disease prediction system using machine learning technique. In 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC) (pp. 689-693). IEEE.

[5] Turanoglu-Bekar, E., Ulutagay, G., & Kantarcı-Savas, S. (2016). Classification of thyroid disease by using data mining models: a comparison of decision tree algorithms. Oxford Journal of Intelligent Decision and Data Sciences, 2, 13-28.

[6] Monaco, F. (2003). Classification of thyroid diseases: suggestions for a revision. The Journal of Clinical Endocrinology & Metabolism, 88(4), 1428- 1432.

[7] Begum, A., & Parkavi, A. (2019, March). Prediction of thyroid disease using data mining techniques. In 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS) (pp. 342-345). IEEE.

[8] Chen, Y., Li, D., Zhang, X., Jin, J., & Shen, Y. (2021). Computer aided diagnosis of thyroid nodules based on the devised small-datasets multi-view ensemble learning. Medical Image Analysis, 67, 101819.

[9] Kousarrizi, M. N., Seiti, F., & Teshnehlab, M. (2012). An experimental comparative study on thyroid disease diagnosis based on feature subset selection and classification. International Journal of Electrical & Computer Sciences IJECS-IJENS, 12(01), 13-20.

[10] Ouyang, F. S., Guo, B. L., Ouyang, L. Z., Liu, Z. W., Lin, S. J., Meng, W., ... & Yang, S. M. (2019). Comparison between linear and nonlinear machine-learning algorithms.

[11] Ahmed, Jamil, and M. Abdul Rehman Soomrani. "TDTD: Thyroid disease type diagnostics." 2016 international conference on intelligent systems engineering (ICISE). IEEE, 2016.

[12] ]Aleksandar, J., Timothy, A. T., & Arye, N. (2004). OFDM Channel Estimation in the Presence of Interference. IEEE Transactions on Signal Processing, 52, 3429-3439.

[13] Ammulu, K., and T. Venugopal. "Thyroid data prediction using data classification algorithm." Int. J. Innov. Res. Sci. Technol 4.2 (2017): 208-212.

[14] Antoine. V., Quost, B., & Masson, H. M., & Denoeux, T. (2012). CECM: constrained evidential c-means algorithm, Computational Statistics and Data Analysis, 56(4), 894–914.

[15] Aswathi, A. K., and Anil Antony. "An intelligent system for thyroid disease classification and diagnosis." 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)