

RESEARCH ARTICLE



A Novel Classification Methodology for Thyroid Cancer using C4.5 with Firefly Optimization Algorithm (CFOA)

 OPEN ACCESS

Received: 09-05-2022

Accepted: 06-06-2022

Published: 14-07-2022

R Vanitha^{1*}, K Perumal²¹ Research Scholar, Department of Computer Applications, Madurai Kamaraj University, Madurai² Professor, Department of Computer Applications, Madurai Kamaraj University, Madurai

Citation: Vanitha R, Perumal K (2022) A Novel Classification Methodology for Thyroid Cancer using C4.5 with Firefly Optimization Algorithm (CFOA). Indian Journal of Science and Technology 15(26): 1324-1335. <https://doi.org/10.17485/IJST/v15i26.996>

* Corresponding author.

vanithachezian2004@gmail.com

Funding: None

Competing Interests: None

Copyright: © 2022 Vanitha & Perumal. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Published By Indian Society for Education and Environment ([iSee](https://www.indjst.org/))

ISSN

Print: 0974-6846

Electronic: 0974-5645

Abstract

Objectives: The main objective of this research is to detect thyroid cancer in its early stages and to improve the accuracy using a novel method C4.5 with Firefly Optimization Algorithm (CFOA). This research also focuses on developing an effective machine learning-based accurate prediction model. **Methods:** The images in this investigation are associated to thyroid disorder with 4672 samples of people including both females and males having hypothyroidism and hyperthyroidism, as well as healthy people without thyroid disorder. The data was gathered for one year with the primary goal of classifying thyroid disease using machine learning (ML) algorithms. These data comprise of Gender, Age, Thyroid hormone (T4), Triiodothyronine (T3), Thyroid stimulating hormone (TSH), etc. The performance of this proposed method was measured using parameters such as accuracy, precision, F1-score and recall. **Findings:** The performance of this proposed method was assessed with the state-of-the-art existing methods like Naive Bayes (NB) algorithm, K-Nearest Neighbor (KNN) Algorithm and Adaboost. The proposed algorithm showed maximum precision of 0.9935, recall of 0.9971, F1-score of 0.9951 and accuracy of 0.9981 respectively when compared with the existing algorithms. **Novelty:** A novel algorithm C4.5 with Firefly Optimization Algorithm was proposed in this paper to speed-up and to increase the effectiveness of the machine learning algorithm. Moreover, this study focuses on developing an accurate prediction model by comparing classification algorithms based on accuracy and confusion matrices and then identifying the most effective classifier based on performance.

Keywords: Thyroid cancer; Classification technique; Hypothyroidism; Hyperthyroidism; Accuracy; Precision; F1score; Recall

1 Introduction

The contribution of this present work is as follows: (a) Enhanced pre-processing is carried out to remove the noise, improve the quality of the image, enhancing some features and so on, (b) Selecting the novel optimal method that is the C4.5 with Firefly

Optimization Algorithm (CFOA) for the proposed work, (c) Comparing the proposed technique with other related techniques in order to prove that the proposed technique is the best technique for diagnosing the thyroid at the early stages.

There have been plenty of processes executed in recent years to detect discrete thyroid disorders. Many researchers have made use of numerous machine learning techniques⁽¹⁾. The authors demonstrated an adequate approach and certainty in identifying diseases similar to thyroid disease through the operation that includes different algorithms and datasets connected with the process that will be done in the forthcoming perspectives to achieve effective and more reasonable results⁽²⁾. The purpose of this article is to interpret numerous techniques of machine learning tools and statistical qualities that have become popular in recent years for the understanding of thyroid disorders with conviction by various authors to achieve various opportunities and techniques⁽³⁾. Machine learning algorithms such as Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), Artificial Neural Network (ANN) and Support Vector Machine among others are widely used in disease prediction^{(4), (5), (6)}.

According to the researchers, the main goal of investigations is to suggest a Machine Learning approach for improving the preciseness of hypothyroidism detection by incorporating patient queries with test outcomes during the detection procedure. The second objective is to decrease the hazards associated with dialysis problems. Data are utilized from UCI database, which contained 3163 samplings. Among them 151 are hypothyroid and the remaining were hypothyroid. The logical determination was resolved if the new instances were hypothyroid. To eliminate the unbalanced distribution, distinct selection approaches were used in data collection, and models were developed to detect hypothyroidism using K-Nearest Neighbor (KNN), Logistic Regression (LR) and Support Vector Machine (SVM) classifiers. In this regard, this article revealed the impact of selection methods on the detection of hypothyroidism. When compared to all the models, the Logistic Regression classifier formed the optimal results. For this examination, the dataset utilizing over-sampling techniques was trained and obtained a precision of 97.8 %, the F-Score of 82.26%, Matthews Correlation Coefficient (MCC) of 81.8 % and ROC of 93.2 %.

New statistical analysis and data mining techniques are utilized by researchers to develop tools that help healthcare professionals to easily and efficiently diagnose thyroid related diseases. Useful knowledge can be extracted from the databases where a significant amount of relevant data is stored. A new decision-based hybrid system for the diagnosis of thyroid diseases is presented here. The proposed system consists of three stages. In the first stage, 25 features of the dataset (retrieved from the University of California Irvin machine learning repository) were reduced using Information Gain method to avoid data redundancy and reduce computation time. In the second stage, the missing values in the dataset are dealt with K-Nearest Neighbor (KNN) weighting pre-processing scheme. Finally, the resultant data is provided as input to the Adaptive Neuro-Fuzzy Inference System for the purpose of input-output mapping in the last stage of this proposed system. Classification accuracy for this proposed approach was calculated to be 99.1%, whereas sensitivity and specificity results were 94.77% and 99.70%, respectively. This approach is able to get highest classification accuracy with minimum possible features of the dataset and can be applied to diagnose other lethal diseases.

For researchers, medical diagnosis and the extraction of designs can convert into applicable knowledge is a difficult task. Medical documents depend on real-time data with increased dimensionality, making pattern extraction even more difficult. The prediction of various diseases, such as thyroid, diabetes, and cancer, is frequently hampered by their high dimensionality. Machine learning is the process of extracting facts and building an understanding base on massive amounts of information that stays useful. These functionalities are accomplished using clustering, classification, prediction and regression. Reducing dimensionality for the purpose of knowledge discovery is an important factor that aids in the prediction/decisiveness for diverse fields such as medical identification, business modelling, and data examination. The article described a suitable method for thyroid classification that combines C4.5 and the Random Forest Classification Technique (CCTML). When the exploratory outcomes are compared to other traditional approaches, the CCTML results show more accurateness⁽⁷⁾.

Research works were carried out research based on the application of a two Multilayer Perceptron (MLP) classifier for categorizing thyroid disorders namely thyroid, hypothyroid and hyperthyroid, as well as achieving maximum accuracy in the shortest amount of time⁽⁸⁾. Two multilayer perceptron classifiers were designed in this work. One classifier is for basic classification, which is used to determine whether the patient is hypothyroid, euthyroid, or hyperthyroid. The classifier has 97.4 percent accuracy and is thus very reliable. If the patient has hypothyroidism, the 5 features are fed into the second classifier, which determines whether the patient has primary, secondary, or complementary hypothyroidism. The second classifier had a classification accuracy of around 91.7 percent.

Recently, authors researched based on an Extreme Machine Learning technique based on Graph Clustering Ant Colony Optimization to detect thyroid threats⁽⁹⁾. This method identified risk factors for thyroid disease. It is critical to identify nodules at an earlier stage. Ultrasonography (US) is a useful mechanism for detecting and distinguishing malignant nodules. The nodules in the US emerge in a variety of heterogenic states that are challenging for physicians to distinguish. Furthermore, the extensive numeral of features obtainable in US characteristics improves calculation time and classification intricacy. The detection of thyroid nodules will improve classification by choosing the best features. The primary objective of this approach

is to distinguish between malignant and benign nodules. Sensitivity, accuracy, AUC and specificity are parameters employed to assess the implementation of both classification and feature selection. According to the experimental results, the suggested method is especially better than the other methods with an accuracy of 98.5 percent. As a result, it is thought to be an effective mechanism for analyzing thyroid nodules with minor complicatedness and less calculation time.

Thyroid disease diagnosis necessitates the accurate performance of operational data from the thyroid gland in order to form hormones that restrain the human body’s metabolic rate⁽¹⁰⁾. Thyroid illnesses are organized based on the number of hormones formed along with hyperthyroidism producing more hormones and hypothyroidism producing lesser than the needed numeral of hormones. Thyroid illness is a major problem in developing nations due to a lack of perception and earlier detection. Over the period of time, the use of machine learning methods serves as an optional technique for the increasing thyroid disorder detection. Here, a novel intelligent hybrid Decision Support System (DSS) is proposed for thyroid disorders that are established on Linear Discriminant Analysis (LDA), Adaptive Neuro-Fuzzy Inference System (ANFIS) and K-Nearest Neighbor (KNN) weighed pre-processing. LDA reduces the dimensionality of the disease from the dataset and destroys excessive features in the first stage of this technique. In the subsequent stage, specified attributes are pre-processed utilizing a KNN-based weighted pre-processor. In the final stage, pre-processed designated features are fed into the ANFIS as an intake for detection. In order to authenticate the prevailing implementation of the system, the suggested approach was tested on datasets obtained from the University of California Irvin’s Machine Learning repository. The calculated classification analysis based on this approach’s accuracy, sensitivity, and specificity values was 98.5 percent, 94.7 percent, and 99.7 percent. This method can also be applied to the detection poisonous diseases in order to achieve ultimate accurateness with the fewest attributes in the dataset. The comparative investigation of various classifiers is described in Table 1.

Table 1. Comparison of various classifiers based on accuracy

Authors	Algorithms	Classification Accuracy
Sparano et al. ⁽⁶⁾	Logistic Regression	97.8%
Dharamkar et al. ⁽⁷⁾	Fusing C4.5 and random forest	97%
Marques et al. ⁽⁸⁾	Multilayer Perceptron	97.4%
Yazdani et al. ⁽¹¹⁾	Graph Clustering Ant Colony Optimization	98.5%
Jha et al. ⁽¹²⁾	Linear discriminant analysis, kNN-based weighed preprocessing, and adaptive neuro fuzzy inference system	98.5%

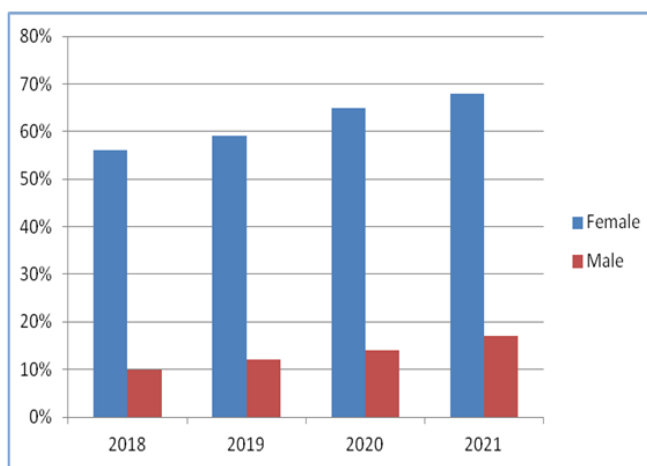


Fig 1. Gender-wise percentage of people affected in India

Figure 1 shows the percentage of people who have thyroid disease in India from 2018-2021.

2 Methodology

2.1 Collection of Data

The images in our investigation are acquired from <https://www.kaggle.com/datasets/yasser hessein/thyroid-disease-data-set> for analyzing and diagnosing diseases, and the kind of data acquired is associated to thyroid disorder with 4672 samples of people including both females and males. These samples include people who have hypothyroidism and hyperthyroidism, as well as healthy people without thyroid disorder. The data was gathered over one year with the primary goal of classifying thyroid disease using ML algorithms. These data comprise of gender, age, Thyroid Hormone (T4), Triiodothyronine (T3), Thyroid Stimulating Hormone (TSH) and so on. As exposed in Table 2, the information obtained comprises of 20 features and all of them were used in the investigation.

Table 2. Description of the Dataset

S. No.	Attribute	Description	Data Type	Value Range
1	Gender	Female or Male	Integer	[0,1]
2	Age	In Years	Real	[0.00,0.90]
3	Pregnant	During pregnancy	Integer	[0,1]
4	Sick	Illness	Integer	[0,1]
5	ATD	Antithyroid drug medication	Integer	[0,1]
6	TSH	Thyroid-stimulating hormone	Real	[0.0,0.53]
7	Class	Category	Integer	(1-3)
8	Li	Lithium	Integer	[0,1]
9	Query thyroxine	-	Integer	[0,1]
10	T3	Triiodothyronine	Real	[0,1]
11	Hyperthyroid	-	Integer	[0,1]
12	Hypothyroid	-	Integer	[0,1]
13	TSH_M	Thyroid stimulating hormone medication	Real	[0,1]
14	TT4	Total Thyroxin	Real	[0.0020, 0.6]
15	T4	Thyroid hormone	Real	[0,1]
16	Thyroid surgery	-	Integer	[0,1]
17	Tumor	Cyst	Integer	[0,1]
18	Hypopituitary	-	Integer	[0,1]
19	T4U	Thyroxin utilization rate	Real	[0.017, 0.233]
20	FT1	Free Thyroxin Index	Real	[0.0020, 0.642]

2.2 Data Pre-processing

Pre-processing is used to determine various kinds of concerns like noisy data, missing data, redundant data or values, etc. High-quality information leads to high-quality outcomes and the cost of ML algorithms are reduced. It is a minor but significant process. Generalizations, cleaning, feature removal, feature selection, and so on are the examples of data processing.

Raw data has lower signal than audio, lost values, and inconsistencies that affect data processing results, partitioning to help improve the data aspect, improve efficiency, and simplicity in the extraction process⁽¹³⁾. Pre-processing refers to a program that processes input data values and also generates output that is used as input to a compiler.

2.3 Classification

The primary goal of ML algorithms is to distinguish between three types of thyroid disorder. The three types of diseases namely hyperthyroidism, hypothyroidism and stable patients with no thyroid problems were examined in this section.

2.3 Decision Tree

Decision Tree is a Supervised Learning (SL) approach that can be utilized to differentiate between Regression problems, but it is most commonly used to solve separation problems. This approach depends on a decision-boosting machine that is examined

for forecasting the energy utilization factor in order to attempt a tree-based process to identify the most effective predictors. The model is designed to improve the decision necessitates including thousands of trees, separately expanded by employing data from the earlier tree. A few tuning parameters are available like the shrinkage parameter, number of trees, total splits in each tree, etc. To construct the learning computation, a split and conquer method, also known as divide and conquer is used⁽¹⁴⁾.

An arrangement of attributes connects the arrangements of instances. The decision tree is made up of leaves and hubs, with leaves representing the class of an illustration that meets the requirements and hubs representing a trial on the measures of a trait. Standards can be obtained by observing the track from the core to the leaf and operating the hubs along the way as pre-conditions for the authority to anticipate the category at the leaf. The tree must be pruned to terminate excessive pre-conditions and duplications^{(11), (15)}.

Decision Trees are based on the Sum of Product (SOP) representation. Each branch from the tree's core to a leaf node along with the category is an intersection (product) of weights and dissimilar branches terminate in that category form a predicate (sum)⁽¹⁶⁾. The foremost challenge in executing a decision tree is deciding which features should be recognized for the root node at every level⁽¹⁷⁾⁻⁽¹⁸⁾. Addressing this issue is directed as an attribute assignment^{(19), (20)}. Different feature selection metrics are utilized at every level to classify the feature that can be regarded as the root node⁽²¹⁻²⁵⁾. Figure 2 represents the structure of the decision tree.

Pseudo Code for Decision Tree

- Step 1: Consider placing the dataset most acceptable feature at the pinnacle of the tree.
- Step 2: Subdivide the training set into smaller subsets.
- Step 3: Replicate steps 1 and 2 per subset until after obtaining the leaf node and all of the tree components.

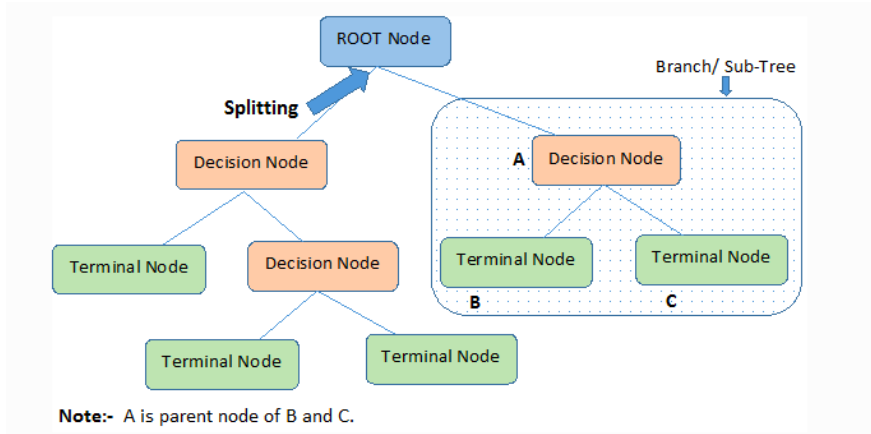


Fig 2. Structure of the Decision Tree

Procedure for creating Decision Tree

- Step 1: Initially, the complete training set is measured as the core (root).
- Step 2: Domain-specific feature weights are preferred. If the weights are successive, they are discretized prior to the model being constructed.
- Step 3: Documents are distributed in a recursive manner depending on feature weights.
- Step 4: A statistical procedure is employed to specify which features should be identified as the tree's core (root) or an interior node.

The architectural process C4.5 decision tree technique was used to perform disease classification techniques. The input data is loaded into databases to perform the classification process. Input data can have missing values, noisy values, or be consistent. Such data must be pre-processed as part of the pre-processing step.

The optimal choice of features is a critical step in improving classification. The best features are chosen, and the dataset is separated into testing and training data for the classification process. The classification algorithm is fed with training data and authorized with testing process. The overall performance of the data is measured using parameters such as accuracy, precision, F1-score and recall. Finally, the prediction values are displayed in the results. The prediction process flow diagram is depicted in Figure 3.

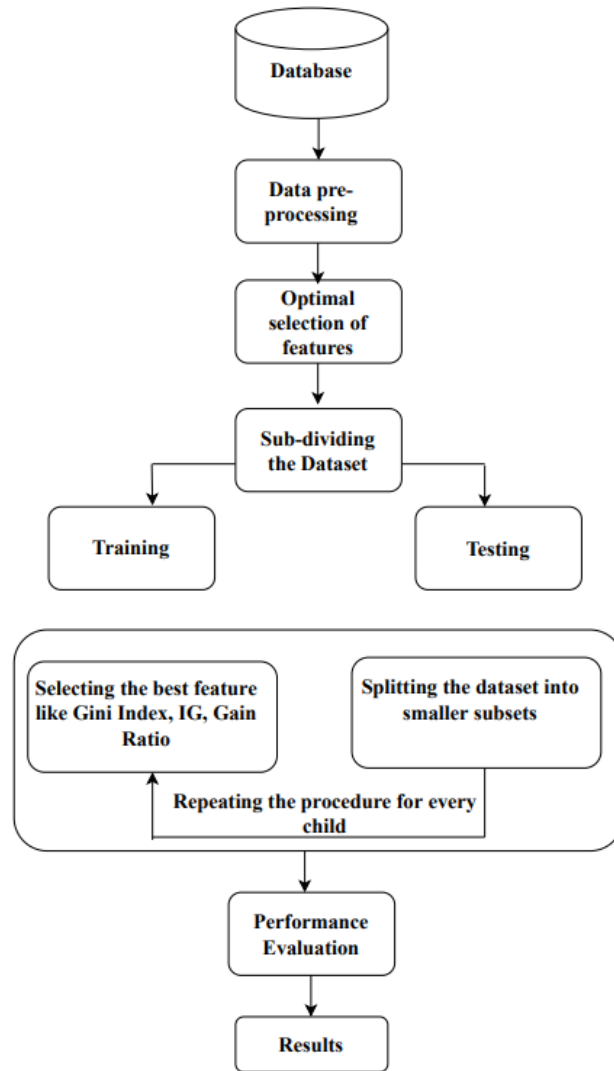


Fig 3. Proposed flowchart with C4.5 with Firefly algorithm

Algorithm 1: C4.5

Input: X_{ab} matrix consisting of continuous parameters
 X_0 : feasible parameter;
 cp_0 : feasible cut point;
 For each parameter $X_i, i=1, \dots, n$ do
 Sort parameter values a_{1i}, \dots, a_{ni}
 Identify all possible cut point $ccp_{1i}, \dots, ccp_{mi}$
 For each cut point $ccp_{ij}, i=1, \dots, m$ do
 Compute information gain $Gain(A, ccp_{ij})$
 Choose the feasible cut point cp_{0i}
 Compute the splitting performance $Split(A, ccp_{0j})$
 Compute gain ratio $Gain_ratio(A, ccp_{0j})$
 Choose the feasible parameter X_0 and cut point cp_0
 End

2.3.2 Optimization model for the classification

Firefly algorithm is a meta-heuristic algorithm based on flashing patterns and behavior of fireflies⁽¹³⁾. Researchers applied the following assumptions regarding fireflies. The firefly algorithm is a heuristic swarm optimization algorithm that was developed to simulate the flashing activities of fireflies and the phenomenon of bioluminescent communication in problem-solving. The Firefly algorithm was revealed to be effective in resolving a wide variety of optimization issues. Because fireflies are unisexual, any firefly will be concerned to any firefly. The more visible firefly will attract the less visible firefly. However, the intensity of the light declines as the length between the fireflies enlarges. If there is no brighter firefly near, the firefly will progress at random. Figure 4 represents the processing of fireflies to obtain the best solution.

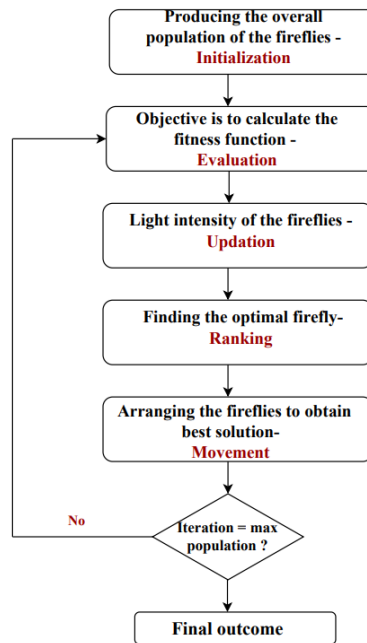


Fig 4. Processing the fireflies for optimal solution

Algorithm 2: Firefly

Input: X(a)

Output: ranking firefly

Procedure: Selecting from the total populace

begin:

X(a), a=(a₁,a₂, a₃,..., a_n)

For total populace m, firefly a_j, where j = 1,2,3,...,m

Compute light intensity l, connected X(a)

while do (k < MaxVal)

define γ

for (x=1;x<n; x++) do

for (y=1;y<n;y++) do

if then l_y > l_x

pull x towards y

compute distance of each firefly

estimate Intensity I

return ranking of firefly

3 Evaluation of the Proposed Method

The proposed C4.5 algorithm extends the ID3 algorithm by determining the best feature based on the Information Gain (IG) ratio. It can manage continuous features by suggesting two separate tests based on the type of attribute value. In order to build the decision tree during the training phase, the C4.5 employs a top-down design depending on the divide-and-conquer methodology. Before producing nodes from the core (root) to the leaves, it locates the training dataset and utilizes the IG ratio as a metric to identify splitting features. Each instructive route from the origin node to the leaf node comprises a decision rule for determining which category a unique instance belongs to.

To account for unknown attribute values, the origin (root node) contains the entire training collection, with all training case values set to 1.0. The algorithm terminates if all of the present node’s training cases belong to a single class. Conversely, if all of the training possibilities are from more than one category, this technique computes the IG ratio for every feature. To split information at the node, the attribute with the greatest IG ratio is chosen. The IG ratio for a discrete feature is calculated by dividing training sets of the existing node in the operation of every value. If the attribute is continuous, a threshold value for splitting must be determined.

The decision tree method is a Machine Learning (ML) technique that employs a persistent data subdividing tool among specified metrics. Interior nodes represent features, branches describe determination laws, and every leaf node interprets the outcome in the recursive DT algorithm. The Top-Down approach was used to categorize the data, which implicates ordering from the root (origin) to a leaf/terminal (last) node. It divides the dataset into small subsets after acquiring the training dataset by using the Information Gain, Gain Index, Gain Ratio and Entropy.

(a) Information Gain (IG)

IG is calculated for splitting by duplicate weight entropies of every branch from the authentic entropy. When using these metrics to qualify a DT, the best splitting process is specified by maximizing the information gain.

$$Information\ Gain = Entropy\ (before) - \sum_{j=0}^k \frac{a_j}{a} Entropy\ (j,\ after) \tag{1}$$

(b) Gain Index (GI)

The Gain Index is computed by deducting the summation of every squared probability.

In the majority of datasets, the Information Gain and Gini Index are calculated. The procedure is replicated for every child row until all rows fit to the exact class and no further features are required. Gini Index is operated to improve detection accurateness and preciseness.

$$gini = 1 - \sum_{n=1}^c (p_i)^2 \tag{2}$$

(c) Entropy

The entropy of a random variable is a measure of its anticipation. It calculates the average portion of the information that is lost when the random variable’s value is unknown.

$$E(S) = \sum_{i=1}^{o^c} - p_i \log_2 p_i \tag{3}$$

(d) Gain Ratio

The gain of information is biased toward selecting features with a huge volume of weights as root nodes. It indicates that chosen features are represented with distinct weights. Determine the information gain of each attribute separately, and then calculate the average information gain. Second, compute the gain ratio of all features whose calculated information gain are higher than or equivalent to the computed average information gain, and then specify the feature with the highest gain ratio to be further subdivided.

$$Gain\ ratio = \frac{Information\ Gain}{Split\ Infor} = \frac{Entropy\ (before) - \sum_{j=0}^k \frac{a_j}{a} Entropy\ (j,\ after)}{\sum_{j=1}^k \frac{a_j}{a} w_j \log_2 w_j} \tag{4}$$

This proposed algorithm uses a top-down greedy search strategy to travel the potential decision-making distance, never retracing and reanalyzing prior choices. This algorithm’s metric is employed for determining the best feature at every point

of the tree expansion referred to as information gain. At the beginning stage, a predictive model must be given proper training utilizing available information.

Later, reliability and accuracy is verified specifically to predict proven test outputs. Various methods were utilized like the Gini index or information gain, to discriminate the optimal feature for splitting nodes utilizing Attribute Selection Measures (ASM). This feature then divides the dataset into more diminutive subsets for until no longer a child additional feature is found. Initially, the model is built using training data. The accuracy of the model is defined, and it is enhanced by evaluating the recognized output by utilizing data. Finally, the prototype can be employed to forecast forthcoming results. After the computational process, the predictive score will be computed. These values serve as a percentage of the accurate predictor in the final node of the eligible model (or leaf). Figure 2 represents the decision tree's working process.

3.1 Implementation

The experiment was carried out in MATLAB R2019a on a Windows 64-bit operating system with an AMD Ryzen 5 4500U 2.38 GHz processor and 8 GB of RAM.

3.1.1 Classification Accuracy (ACC)

Accuracy is measured in terms of positives and negatives, with a scale ranging from 0 to 100 percent.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (5)$$

Here, TN denotes True negative, FN denotes False Negative, TP denotes True Positive, and FP denotes False Positive values.

3.1.2 Precision

Precision and recall seek to measure the percentage of True Negative (TN) and True Positive (TP). Precision is the capacity of a classifier not to recognize a positive specific case that is essentially negative, and it is stated as follows:

$$Precision = \frac{TP}{(TP + FP)} \quad (6)$$

3.1.3 Recall

Conversely, recall assesses the model's sensitivity. It is mentioned as the ratio of exactly predicted value for a class to the sum of patients occurred, and it is calculated as follows:

$$Recall = \frac{TP}{(TP + FN)} \quad (7)$$

3.1.4 F1 Score

Many real-world classification problems have unequal class distributions. When the TP and TN seems to be more significant, accuracy is utilized, but F1-score is employed when the FN and FP are extremely crucial. As a result, F1-score may be a better metric to use to evaluate our model. The numerical average of precision and recall is considered as F1-Score.

$$F1 - score = 2 \frac{(Precision \times recall)}{(Precision + recall)} \quad (8)$$

Maximum of the available datasets were found to be suffering from inappropriate as well as noise features which might result the classification procedure to generate disagreeable outcomes, consequently in this stage feature selection procedure was employed for reducing the dimensionalities of dataset as well as for eliminating unwanted features.

Concerning machine learning scenario, extraction or selection of finest feature possesses significant part in order to improve the entire classification procedure. For investigating the prominence of features selection concerning classification, dissimilar datasets were classified deprived of features selection.

4 Results and Discussions

The results clearly showed that, feature selection was considered to be one of the significant stages in classification, in which the dataset suffers from noise as well as inappropriate data that might have negative effect over classification accuracy. Consequently,

outcomes with no feature selection processes may be poorer because of dimensionalities as well as noise data, however accounting with feature selection, overall accuracy was enhanced.

Thyroid patients were divided into three groups: normal, hyperthyroidism, and hypothyroidism. There were 4672 samples after data pre-processing, 3066 images were taken for training and 1606 images were taken for testing out of the total.

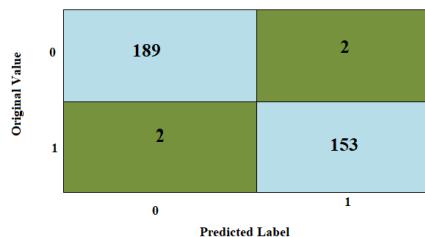


Fig 5. Confusion Matrix

They are split into 3574 training and 1098 testing samples, with a train and test split ratio of 70:30. A confusion matrix is computed to alleviate the problem with FP and FN numbers. Figure 5 displays the confusion matrix along with the performance measures of the deployed classifier. Figure 6 depicts the analysis of the different classifiers based on the selected metrics.

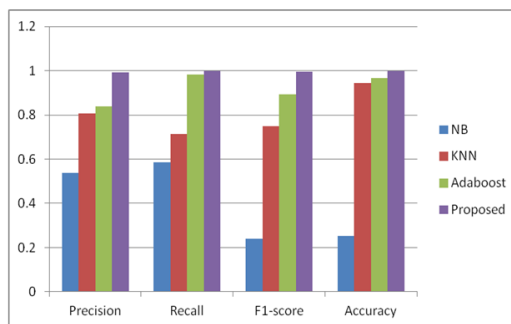


Fig 6. Comparative analysis of the various parameters of different classifiers

Naive Bayes (NB) represents a simpler classification procedure to predict the model possessing perfect semantics that represents probabilistic learning technique on the basis of Bayesian theorems. NB classifiers assume the values of a single attribute to be not depending over values of other attributes, thereby NB adopts that the existence or non-existence of specific attributes in prediction possessing maximum probabilities. In case of this proposed method, the collaborative filtering helps in the construction of recommendation systems using machine learning methods for filtering hidden data as well as for predicting that users are seeking a given feature.

K-Nearest Neighbors (KNN) techniques does not make any assumption regarding distribution of information on which it was grounded. Normally, KNN models were found to be successfully employed over specific datasets. KNN is advantageous while functioning over few real-world datasets. In this proposed research, the datasets were associated with thyroid disorder possessing hypothyroidism and hyperthyroidism, moreover fit people without thyroid disorder. KNN makes delay in testing as well as requires greater amount of timing as well as memory space. In KNN, K specifies the control parameter in the prediction models, however, KNN shall not be learning instantaneously from training sets but it upholds the datasets and employs them for classification, which is not the case in this proposed method.

AdaBoost machine learning models characteristically encompasses several thousands of shallower decision tree, possibly causing much larger search spaces. Considering the input $x \in X$, it becomes possible for reducing these spaces exponentially on taking into account only the decision path of x in all the decision trees as well as dropping every additional branches. These paths preserve every information regarding the determination of $g(x)$. In this proposed method, a statistical procedure was employed to identify the features to be identified as tree’s root or interior nodes.

ompares the accuracy obtained with the deployed classifiers. The proposed method achieved the highest accuracy of 99.81 percent, followed by Adaboost, KNN, and NB, which achieved 96.65 percent, 94.47 percent, and 25.09 percent, respectively.

Table 3. Time calculation based on seconds

Classifier	Speed (in terms of seconds)
NB	0.0923
KNN	0.5453
Adaboost	4.1251
Proposed	1.117

When the TP and TN are more important, accuracy is used, but F1-score is used when the FN and FP are critical. When an F1-score is 1, it is considered perfect. Among all the classifiers, the suggested method has the highest F1-score of 0.9951. Thus, in terms of precision, accuracy, F1-score and recall, the proposed method outperformed the other models. This demonstrates that the proposed classifier is significantly better at Thyroid disease prediction.

In real-time applications, it is required to prioritize model accuracy over speed. Results shows that while NB and KNN take less time to train, their accuracy is very low. Adaboost accepted 4.1251s in our case, which really is slower than the most precise proposed model. In this investigation, the proposed method is more suitable for further processing.

5 Conclusion and Future works

In this article, a novel algorithm C4.5 with Firefly Optimization Algorithm (CFOA) was proposed to speed-up and enhances the effectiveness of the machine learning algorithm. The speed of the NB, KNN, Adaboost and proposed algorithms are 0.0923 sec, 0.5453 sec, 4.1251 sec and 1.117 sec respectively, that clearly demonstrates that the proposed classifier is suggestively faster in thyroid disease prediction. Moreover, the accuracy is 0.2509, 0.9447, 0.9665 and 0.9981 for NB, KNN, Adaboost and proposed algorithms respectively, that shows that the proposed algorithm displayed highest accuracy of 99.81%

In future, more additional attributes necessitate prominent clinical tests, which are both costly and time-consuming. As a result, there is a necessity to design such algorithms and predictive standards that demand the least number of metrics from a person to analyze thyroid disease and save both money and period of time for the patient.

References

- 1) Vasile CM, Udriștoiu AL, Ghenea AE, Popescu M, Gheonea C, Niculescu CE, et al. Intelligent Diagnosis of Thyroid Ultrasound Imaging Using an Ensemble of Deep Learning Methods. *Medicina*. 2021;57(4):395–395. Available from: <https://doi.org/10.3390/medicina57040395>.
- 2) Hu M, Asami C, Iwakura H, Nakajima Y, Sema R, Kikuchi T, et al. Development and preliminary validation of a machine learning system for thyroid dysfunction diagnosis based on routine laboratory tests. *Communications Medicine*. 2022;2(1). Available from: <https://doi.org/10.1038/s43856-022-00071-1>.
- 3) Mirbabaie M, Stieglitz S, Frick NRJ. Artificial intelligence in disease diagnostics: A critical review and classification on the current state of research guiding future direction. *Health and Technology*. 2021;11(4):693–731. Available from: <https://doi.org/10.1007/s12553-021-00555-5>.
- 4) Zhang X, Lee VCS, Rong J, Liu F, Kong H. Multi-channel convolutional neural network architectures for thyroid cancer detection. *PLOS ONE*. 2022;17(1):e0262128–e0262128. Available from: <https://doi.org/10.1371/journal.pone.0262128>.
- 5) Yadav DC, Pal S. Performance based Evaluation of Algorithmson Chronic Kidney Disease using Hybrid Ensemble Model in Machine Learning. *Biomedical and Pharmacology Journal*. 2021;14(3):1633–1645. Available from: <https://dx.doi.org/10.13005/bpj/2264>.
- 6) Sparano C, Verdiani V, Pupilli C, Perigli G, Badii B, Vania Vezzosi, et al. Choosing the best algorithm among five thyroid nodule ultrasound scores: from performance to cytology sparing—a single-center retrospective study in a large cohort. *European Radiology*. 2021;31:5689–5698. Available from: <https://doi.org/10.1007/s00330-021-07703-5>.
- 7) Dharamkar B, Saurabh P, Prasad R, Mewada P. An Ensemble Approach for Classification of Thyroid Using Machine Learning. In: Das H, Pattnaik P, Rautaray S, and LK, editors. *Advances in Intelligent Systems and Computing*; vol. 1119. Springer Singapore. 2020;p. 13–22. Available from: https://doi.org/10.1007/978-981-15-2414-1_2.
- 8) Marques G, Ferreras A, Torre-Diez IDL. An ensemble-based approach for automated medical diagnosis of malaria using EfficientNet. *Multimedia Tools and Applications*. 2022. Available from: <https://doi.org/10.1007/s11042-022-12624-6>.
- 9) Zhang X, Lee VCS, Rong J, Liu F, Kong H. Multi-channel convolutional neural network architectures for thyroid cancer detection. *PLOS ONE*. 2022;17(1):e0262128–e0262128. Available from: <https://doi.org/10.1371/journal.pone.0262128>.
- 10) Cordes M, Götz TI, Lang EW, Coerper S, Kuwert T, Schmidkonz C. Advanced thyroid carcinomas: neural network analysis of ultrasonographic characteristics. *Thyroid Research*. 2021;14(1):16–16. Available from: <https://doi.org/10.1186/s13044-021-00107-z>.
- 11) Yazdani A, Varathan KD, Chiam YK, Malik AW, Ahmad WAW. A novel approach for heart disease prediction using strength scores with significant predictors. *BMC Medical Informatics and Decision Making*. 2021;21(1). Available from: <https://doi.org/10.1186/s12911-021-01527-5>.
- 12) Jha R, Vandana Bhattacharjee, Mustafi A. Increasing the Prediction Accuracy for Thyroid Disease: A Step Towards Better Health for Society. *Wireless Personal Communications*. 2022;122(2):1921–1938. Available from: <https://doi.org/10.1007/s11277-021-08974-3>.
- 13) Nayak J, Naik B, Dinesh P, Vakula K, Dash PB. Firefly Algorithm in Biomedical and Health Care: Advances, Issues and Challenges. *SN Computer Science*. 2020;1(6):311–311.
- 14) Rehman HAU, Lin CYY, Mushtaq Z, Su SF. Performance Analysis of Machine Learning Algorithms for Thyroid Disease. *Arabian Journal for Science and Engineering*. 2021;46(10):9437–9449. Available from: <https://doi.org/10.1007/s13369-020-05206-x>.

- 15) Capitoli G, Piga I, Limperio V, Clerici F, Leni D, Garancini M, et al. Cytomolecular Classification of Thyroid Nodules Using Fine-Needle Washes Aspiration Biopsies. *International Journal of Molecular Sciences*. 2022;23(8):4156–4156. Available from: <https://doi.org/10.3390/ijms23084156>.
- 16) Chandio JA, Mallah GA, Shaikh NA. Decision Support System for Classification Medullary Thyroid Cancer. *IEEE Access*. 2020;8:145216–145226. Available from: <https://doi.org/10.1109/ACCESS.2020.3014863>.
- 17) Islam SS, Haque MS, Miah MSU, Sarwar TB, Nugraha R. Application of machine learning algorithms to predict the thyroid disease risk: an experimental comparative study. *PeerJ Computer Science*. 2022;8:e898–e898. Available from: <http://doi.org/10.7717/peerj-cs.898>.
- 18) Kumar Y, Gupta S, Singla R, Hu YC. A Systematic Review of Artificial Intelligence Techniques in Cancer Prediction and Diagnosis. *Archives of Computational Methods in Engineering*. 2022;29(4):2043–2070. Available from: <https://doi.org/10.1007/s11831-021-09648-w>.
- 19) Mourad M, Moubayed S, Dezube A, Mourad Y, Park K, Torreblanca-Zanca A, et al. Machine Learning and Feature Selection Applied to SEER Data to Reliably Assess Thyroid Cancer Prognosis. *Scientific Reports*. 2020;10(1):5176–5176. Available from: <https://doi.org/10.1038/s41598-020-62023-w>.
- 20) Li H, Weng J, Shi Y, Gu W, Mao Y, Wang Y, et al. An improved deep learning approach for detection of thyroid papillary cancer in ultrasound images. *Scientific Reports*. 2018;8(1):6600–6600. Available from: <https://doi.org/10.1038/s41598-018-25005-7>.
- 21) Nguyen DT, Pham TD, Batchuluun G, Yoon HS, Park KR. Artificial Intelligence-Based Thyroid Nodule Classification Using Information from Spatial and Frequency Domains. *Journal of Clinical Medicine*. 1976;8(11):1976–1976. Available from: <https://doi.org/10.3390/jcm8111976>.
- 22) Layek K, Basak B, Samanta S, Maity SP, Barui A. Stiffness prediction on elastography images and neuro-fuzzy based segmentation for thyroid cancer detection. *Applied Optics*. 2022;61(1):49–49. Available from: <https://doi.org/10.1364/AO.445226>.
- 23) Ajilisa OA, Jagathyraj VP, Sabu MK. Computer-Aided Diagnosis of Thyroid Nodule from Ultrasound Images Using Transfer Learning from Deep Convolutional Neural Network Models. *2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA)*. 2020;p. 237–241. Available from: <https://doi.org/10.1109/ACCTHPA49271.2020.9213210>.
- 24) Rehman HAU, Lin CY, Mushtaq Z, Su SF. Performance Analysis of Machine Learning Algorithms for Thyroid Disease. *Arabian Journal for Science and Engineering*. 2021;46(10):9437–9449. Available from: <https://doi.org/10.1007/s13369-020-05206-x>.
- 25) Cruz JA, Wishart DS. Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Informatics*. 2006;2:117693510600200–117693510600200. Available from: <https://doi.org/10.1177/117693510600200030>.