IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11341

Expert System Based on Multi-Stage Approach Combining Feature Selection with Machine Learning Techniques for Diagnosis of Thyroid Disease

Dr. Avijit Kumar Chaudhuri[1], Shulekha Das[2]

[1]Assistant Professor, Computer Science and Engineering, Techno Engineering College Banipur, Kolkata, India

[2]Assistant Professor, Computer Science and Engineering, Techno Engineering College Banipur, Kolkata, India

Abstract: The thyroid gland produces thyroid hormones levothyroxine (abbreviated T4) and triiodothyronine (abbreviated T3). These hormones play an important role in protein synthesis, body temperature regulation, and total energy generation and regulation. Many disorders affect the thyroid gland, some of which are very frequent, such as hypothyroidism and hyperthyroidism. Thyroid disorders (TD) impact 42 million individuals in India, with hypothyroidism being the most common, affecting one in every ten adults. According to a study report published in the journal Lancet in February 2022[1] type 1 diabetes among people under the age of 25 accounted for at least 73.7% of the overall 16,300 diabetes fatalities in this age group in 2019. This is even though this illness is largely treatable. To reduce such TD, early detection of the disease is essential. A fast, accurate, and interpretable machine learning model is a research subject. Fewer features reduce the computational effort and improve interpretation. A 3-Stage hybrid feature selection approach and several classification models are evaluated on the TD dataset obtained from the kaggle.com website with 29 features and one outcome variable. Stage-1 uses a Genetic Algorithm and Logistic Regression Architecture for Feature Selection and selects 13 features well correlated with the class but not among themselves. Stage-2 utilizes the same Genetic Algorithm and Logistic Regression Architecture for Feature Selection to select 11 features. In Stage-3, Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), Extra Trees (ET), Random Forest (RF), and Gradient Boosting (GDB) are used with the 11 features to identify patients with or without TD. Data splitting, several metrics, and statistical tests are used, along with 10-fold cross-validation, to do a comparative analysis. LR, NB, SVM, ET, RF, and GDB demonstrate improvement across performance measures by reducing the number of features to 11. When compared to prior research, many performance metrics such as accuracy, sensitivity, specificity, f-measure, AUC values, and kappa statistics showed superior outcomes with fewer features. Finally, with 100% classification results, the proposed ensemble model demonstrated its worth. The output findings were compared to those of previous research on the same dataset, and the proposed model was determined to be the most successful across all performance dimensions.

1. https://www.downtoearth.org.in/news/health/1-in-10-indians-have-hypothyroidism-61693

Keywords: Thyroid Disorders, Machine Learning Classifiers, Feature Selection, Genetic Algorithm(GA), Extra Trees(ET), Gradient Boosting(GDB), Random Forest(RF)

1. INTRODUCTION

According to the study[1], one in every 38000 people worldwide has continuous congenital hypothyroidism. TD affects approximately 42 million people worldwide, with the majority living in low-income countries such as India. This ratio appears to be more common among Indians in Mumbai, where it is one in 2640. Over 25,000 hospitals worldwide are already collecting patient data in various formats. TD is common in today's world, and it frequently results in serious damage to life and the body.

The thyroid is a large gland that resembles a butterfly in form. It is located in the bottom portion of the neck and aids in body metabolic management [2]. Thyroid hormones levothyroxine (abbreviated T4) and triiodothyronine (abbreviated T3) are produced by this gland [3, 4]. These hormones play an important role in protein synthesis, body temperature regulation, and total energy generation and regulation [5, 6]. Many disorders affect the thyroid gland, some of which are very frequent, such as hypothyroidism and hyperthyroidism [4]. Hypothyroidism is caused by a lack of thyroid hormone secretion, whereas hyperthyroidism is caused by an excess of thyroid hormone secretion [3,7]. The first instance is hypothyroidism, which is characterized by a deficit or underproduction of thyroid hormones. The symptoms of this condition include weight gain, swelling in the front of the neck, and a low pulse rate, whereas hyperthyroidism

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11341

refers to the thyroid gland producing an excessive amount of thyroid hormone, resulting in elevated blood pressure and pulse rate while having a lower body weight [7, 8]. Blood tests, which may detect TSH, T3, and T4 levels, are a typical way of identifying thyroid problems [9, 10]. In the medical field, the health care industry generates a vast amount of complicated data that is difficult to handle [6]. A variety of machine learning algorithms have lately been applied to investigate and detect various sorts of illnesses. Classical analysis and statistical assessments are two traditional approaches for clinical and medical studies. Researchers employ a variety of classification approaches, including Bayesian networks (BN), SVM, Artificial neural networks(ANN), DT, NB, K-Nearest Neighbour (KNN), and many more [10–13]. Doctors typically interpret a patient's present diagnostic test results to diagnose TD. The diagnosis becomes difficult when the outcomes of tests conflict with each other. In the absence of any known risk factors, the diagnosis is especially challenging, and consistency of results among doctors is difficult to achieve. Human determination of the risks, or a diagnosis of TD based on risk factors, is difficult[14].

Given our improved ability to collect, store, and interpret data to reveal patterns and provide insights, the machine-learning technique is advantageous in this case. Machine-learning approaches have the potential to detect risks early in the course of TD. Any error in disease prediction can be disastrous, and Type II errors are particularly severe. In addition to accuracy, other performance indicators such as consistency, sensitivity, and specificity are important in such research. Several machine-learning techniques used to predict disease in general, and TD in particular, falls short of these additional performance criteria.

This research work presents a multi-stage method to TD diagnosis that combines feature selection with machine learning approaches for increased prediction reliability and accuracy. It is compared against many cutting-edge machine-learning classifiers (LR, NB, DT, SVM with Radial Basis Function kernel, and RF) as well as earlier research on the same data set. We base the comparison on many performance measures and statistical tests (accuracy, sensitivity, specificity, ROC, AUC, Kappa statistic) on 50–50 %, 66–34 %, 80–20 %, and 10-fold cross-validation splits of training and testing data. The goal is to find answers to the following research questions:

Research Question 1: Is the proposed multi-stage technique for TD prediction recommended?

Research Question 2: Does the suggested strategy fulfil the extra criterion of Sensitivity, Specificity, F-Measure, ROC-AUC and Kappa Statistics?

Research Question 3: Is the suggested approach consistent and statistically significant throughout the dataset's various levels of Training and Testing splits?

The flowchart and architectural design of the experimental design and model building is depicted in Figure 1, Figure 2 and Figure 3.

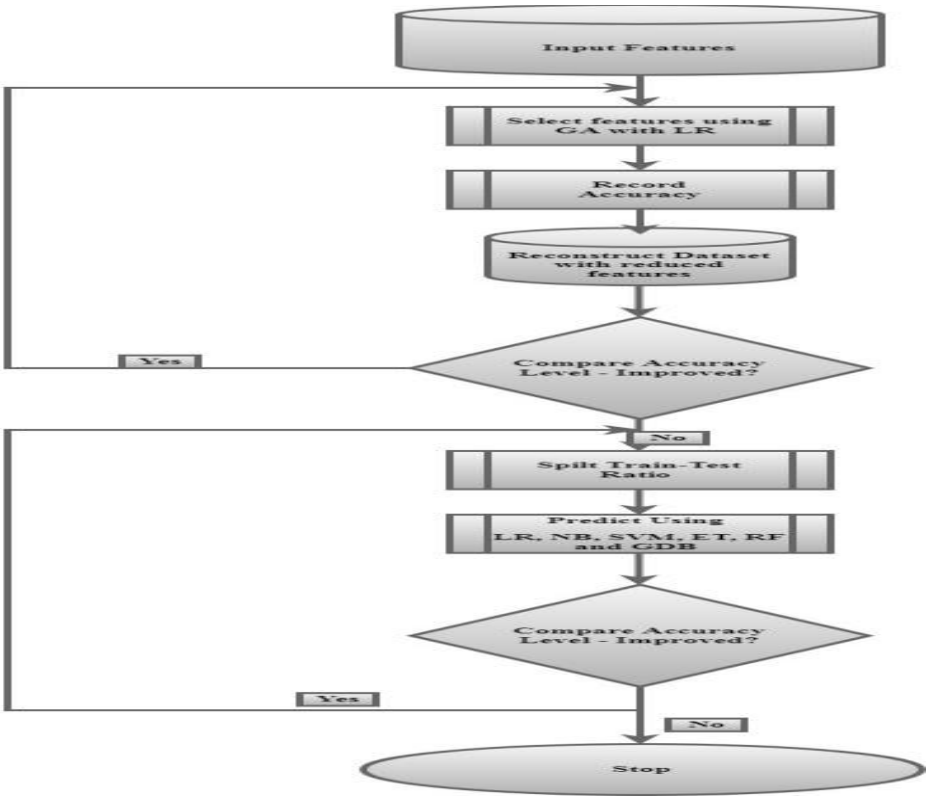


Fig.1. Hybrid Feature Selection and Classification Model

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11341

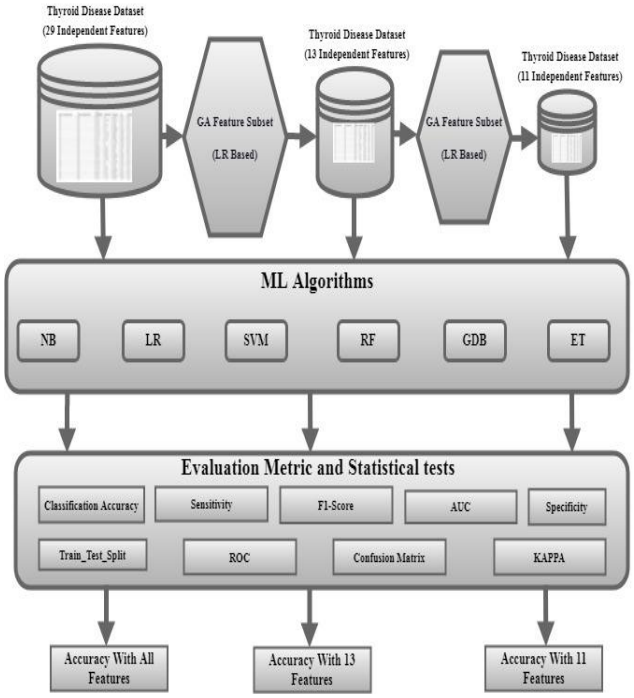


Fig.2. The Architecture for the Proposed System

2.CHOICE OF DATA MINING MODELS

This study investigated data-mining models, specifically LR, NB, SVM, ET, RF, and GDB, for assessment of the reasons of TD and accurate disease prediction. The algorithmic approach depicted in Figure 1 is explained in detail in Figure 3 below.

Step 1: (Dataset Acquisition) All records from the TD dataset are collected and read.
Step 2: (Classify hypothyroid and no- hypothyroid without feature selection) Classification algorithms, namely LR, NB, SVM, RF, ET, and GDB are used to measure the classification accuracy of the TD patients.
Step 3: (Selection of relevant features) GA with LR is applied to get the relevant features.
Step 4: (Classify hypothyroid and no- hypothyroid) Classification of optimal feature subset using algorithms, such as LR, NB, SVM, RF, ET, and GDB and accuracy measured.
Step 5: (Selection of more relevant features) GA with LR is applied once again to get more relevant features.
Step 6: (Classify hypothyroid and no- hypothyroid) Classification of optimal feature subset using algorithms, such as LR, NB, SVM, RF, ET, and GDB and accuracy measured.
Step 7: (Validation) The classifiers are trained using the validation set. A train- test partitioning and 10 - fold cross- validation technique are used for testing purposes.
Step 8: (Performance parameter computation) Computation of various correctness parameters namely accuracy, sensitivity, specificity, f 1 - score, ROC- AUC, and Kappa score.

Fig.3. Hybrid Feature Selection and Classification Model Algorithm

The purpose of feature selection is to identify feature combinations that results the best predictive model for early detection and progression of TD. A GA is used to choose one or more sets of features, and the LR method is used to create a prediction model. The algorithmic approach that integrated GA and LR to predict TD status is depicted in Figure 1. The GA output features are fed into LR, and the resulting variable sets are used by the GA to refine and select the optimal set of features. The next sections go through the criteria for selecting Data Mining Techniques

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11341

2.1 Logistic Regression (LR)

The impact of numerous explanatory variables on a response variable may be studied simultaneously using LR. It returns the linear combination of factors that predicts the likelihood of an event, such as the occurrence of TD or the absence of TD. As a result, it calculates the contribution of a collection of attributes to a binary outcome. It is one of the most often used models, particularly in clinical practice[15], since it allows for the regression of dependent variables over a wide range of variables. Since the 1990s, it has seen a growth in its application in medical research. Its interpretation is simple, and it benefits in speedy decision-making. However, the issue with LR is that it has a propensity to produce over-fitted models[16].

2.2 Naive Bayes (NB)

The NB algorithm calculates a series of probabilities by counting the frequency and value combinations of a given dataset[17]. It is known as a simple probabilistic classifier. The Bayes' algorithm employs conditional probability to determine the probability of a randomly selected feature selected as a classifier on a particular category. NB assumes that any two randomly chosen features are statistically independent of each other, avoiding the problem from a large number of vectors in the Bayesian classifier [18]. This conditional presumption of independence seldom holds in real-world applications; hence, the characterization of this algorithm as "naive." The algorithmic rule, however, performs well and quickly learns numerous supervised classification problems[19]. The Bayes' theorem is one of the most common classification techniques applied to small datasets due to its simplicity, robustness, and accuracy. The performance of the NB classifier with large datasets and datasets with complex attribute dependencies is poor.

2.3 Support-Vector Machine (SVM)

Vapnik created the SVM algorithm as a regulatory algorithm in 1995. The basis of this technique is to use exactness to generalize errors. This technique generates one 'hyperplane' and divides the data into classes to categorize the samples. SVM shows good results for multi-domain or binomial applications in a big data environment[18]. It performs faster after training[20]. However, this method is mathematically complicated and computationally expensive. A large dataset is likely to contain noise, and SVM yields poor results in such cases. This low performance is because SVM makes use of hyperplanes and support vectors that classify in higher-dimensional space. Studies have overcome this drawback by combining SVM with other machine-learning techniques. The efficiency of SVM lies in its use of the appropriate kernel function and fine-tuning [21].

2.4 Random Forest (RF)

RF, a supervised machine-learning algorithm, is a blended arrangement technique based on the statistical learning hypothesis [21, 22]. RF creates multiple DTs and combines them to give the best classification [21, 23]. To generate the individual classifiers, it uses either a bagging or random selection of features. It uses a classifier strategy, called the unweighted majority of class votes, to minimize errors. A large number of trees make an RF from the selected samples. Each tree votes and the most popular class get chosen as the outcome in a classification problem. The introduction of the right kind of randomness impacts the accuracy of RF. The generation of the tree with minimal depth has an advantage as it is independent of how prediction error is measured[24, 25]

2.5 Extra Trees

The 'Extra Trees' Classifier (ETC) is a random forest variant, also known as the 'Extremely Randomized Trees' Classifier. Unlike a random forest, the entire sample is used at each stage, and the decision boundary is chosen at random rather than based on the best results. In practice, the yield is equivalent to, if not better than, a conventional, random forest [26, 27]. It is a collection of decision trees that can be combined with other decision tree techniques such as bootstrap aggregation (bagging) and RF. The ET method creates a large number of un-pruned decision trees from the training dataset. Forecasts are generated in the case of regression by combining the forecasts of decision trees or, in the case of classification, by using the majority vote[14].

2.6 Gradient Boosting (GDB)

GDB is a classification and regression machine learning algorithm. This builds a predictive model in the form of an ensemble of weak classifiers. The application of feature engineering or the direct implementation of boosting algorithms is two ways to improve the accuracy of classification procedures. Friedman [28] introduced the GDB method, which has since been widely used in a variety of clinical settings [29-31]. GDB is a non-parametric, supervised machine learning method. It approximates an indeterminate functional mapping from input explanatory variables to subsequent output variables.

GDB requires three components: (a) an optimized loss feature, (b) a weak learner for prediction, and (c) an additive model to reduce loss functions by adding weak learners.

GDB is a greedy algorithm that overfits a training dataset easily. Regularization approaches that penalize many parts of

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11341

the algorithm and often boost algorithm efficiency by decreasing overfitting [14] may be beneficial.

3.PERFORMANCE METRICS

The number of TD patients classified as thyroid patients is True Positive (TP). False Positive (FP) is the number of non-thyroid patients classified as TD patients. True Negative (TN) is the number of patients that are classified as non-thyroid patients without TD. False-negative (FN) is the number of patients classified as TD patients without TD.

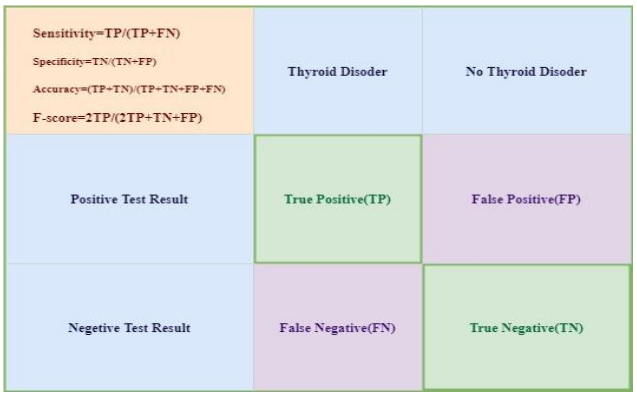


Fig.3. Confusion matrix, performance evaluation metrics and statistical tests

3.1 Sensitivity

The classification function in statistics is a statistical measure of the performance of a binary classification test. The proportion of true positives correctly identified as such is known as sensitivity as illustrated in Figure 3. It is also known as the true positive rate or, in some cases, the recall rate (e.g. the percentage of sick people who are correctly identified as having the condition).[25,32]

3.2 Specificity

It (also known as the true negative rate as shown in Figure 3) is a metric that quantifies the percentage of negatives that are accurately classified(e.g. the percentage of healthy people who are correctly identified as not having the condition). The ideas of type I and type II errors are strongly connected to these two measurements.[18, 25, 32]

3.3 F- Measure

It is critical is defined as the weighted harmonic mean of accuracy and recall. It is critical to compute the F-score because this is a statistic of averages, the best F-score was chosen. Overtraining was avoided at all costs because it might have a detrimental effect on some algorithms. Overtraining is indicated if the training is more accurate than the forecast. A model is deemed excellent if its value is one or it has a low false positive or false negative rate, but a value of 0 indicates poor performance[25]. The F1-score equation is shown in Figure 3.

3.4 Kappa Test

Calculating kappa is used to evaluate validity in terms of sensitivity, specificity, and reliability. Kappa can be used to calculate agreement between two raters (expert analysis and data mining techniques analysis) when classifying the same set of instances. Cohen's kappa defines agreement as the ratio of the percentage of agreement minus the chance agreement to the highest feasible non-chance agreement. As a result, this metric considers classes that may coincide by chance. The percentage of matches in each class determines the probability agreement, which decreases as the number of classes' increases. According to the preceding criteria, a kappa value of 1 indicates complete agreement, whereas a kappa value of 0 indicates agreement is no better than chance[18, 25, 32]. The equation to calculate kappa value is shown in equation (1) below.

$$\text{Kappa Test} = \frac{\text{Observed Agreement} - \text{Expected Agreement}}{100 - \text{Expected Agreement}}$$

where, Observed Agreement = %(Overall Accuracy)

Expected Agreement = (%(TP + FP)× %(TP + FN)) + (%(FN + TN)× %(FP + TN)) (1)

3.5 Area under the curve (AUC)

Receiver operating characteristic (ROC) is plotted between Sensitivity and (1-Specificity). The area under the curve (AUC) measures the degree to which the curve is up in the northwest corner[25, 32].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11341

4. DATASET DESCRIPTION

The data set "Hypothyroid Disease Data Set" was gathered for this research work from https://www.kaggle.com/yasserhessein/thyroid-disease-data-set/version/1?select=hypothyroid.csv to investigate the working principle of the 3-Stage Hybrid Feature Selection methodology and several Classification Models. There were a total of 3772 individuals in the sample, with 3481 (92.3%) hypothyroid and 291 (7.7%) negative. Table 1 lists the features included in the dataset.

Table.1. Description of the TD Dataset

Sl. No.	Variable	Variable Description
1	Age	Integer
2	Sex	Male(1), Female(0)
3	On thyroxine	False(0), True(1)
4	Query on thyroxine	False(0), True(1)
5	On antythyroid	False(0), True(1)
6	Sick	False(0), True(1)
7	Pregnant	False(0), True(1)
8	Thyroid surgery	False(0), True(1)
9	T131 treatment	False(0), True(1)
10	Query Hypothyroid	False(0), True(1)
11	Query Hyperthyroid	False(0), True(1)
12	Lithium	False(0), True(1)
13	Goiter	False(0), True(1)
14	Tumor	False(0), True(1)
15	Hypopitutory	False(0), True(1)
16	Psych	False(0), True(1)
17	Tsh measured	False(0), True(1)
18	TSH	Real
19	T3 measured	False(0), True(1)
20	T3	Real
21	TT4 measured	False(0), True(1)
22	TT4	Real
23	T4U measured	False(0), True(1)
24	T4U	Real
25	FTI Measured	False(0), True(1)
26	FTI	Real
27	TBG Measured	False(0), True(1)
28	TBG	Real
29	Referal source	SVHC, other, SVI, STMW, SVHD
30	Class	negative, hypothyroid

5. RESULTS AND DISCUSSION

This research article addresses the following research questions: Which Data Mining Technique (DMT) is most effective in predicting disorders like thyroid disease? & Which DMT framework may help achieve the three criteria of consistency, sensitivity, and specificity? The authors evaluate the most common methodologies and investigate their ensemble to reach the maximum levels of accuracy, sensitivity, and specificity. Previous researchers have only concentrated on lowering variables to enhance prediction. However, this procedure leads to data loss. Thus, the authors suggest in this research a framework for applying data mining methodologies, measuring consistency with kappa

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11341

statistics, and improving specificity and sensitivity parameters with an ensemble learning approach. As a result, the methodology proposed in this research adds to the well-being of humanity by allowing for better disease prediction.

As shown in Table 1, a 3-Stage Hybrid feature selection strategy and Classification models are assessed using the TD dataset collected from the kaggle.com website, which contains 29 features and one outcome variable. Stage-1 picks 13 characteristics that are strongly connected with the class but not among themselves using a Genetic Algorithm and Logistic Regression Architecture, as shown in Table 2. Stage-2 selects 11 characteristics using the same Genetic Algorithm and Logistic Regression Architecture as shown in Table 3.

Table.2. Thyroid Dataset with 13 Features and 1 Target Variable

Attributes
age, on thyroxine, thyroid surgery, query hyperthyroid, lithium, tumor, psych, TSH measured, TSH, T3, T4U measured, FTI measured, TBG measured, binaryClass

Table.3. Thyroid Dataset with11 Features and 1 Target Variable

Attributes
age, on thyroxine, thyroid surgery, query hyperthyroid, tumor, psych, TSH measured, TSH, T3, T4U measured, FTI measured, binaryClass

Table.4. Comparison of Accuracies with 27, 13 and 11 Features

Train – Test Split	Number of Features	LR	NB	SVM	ET	RF	GDB
50-50	29	0.95	0.23	0.96	0.99	0.99	0.99
	13	0.95	0.36	0.96	0.99	0.99	0.99
	11	0.95	0.36	0.96	0.99	0.99	0.99
66-34	29	0.96	0.22	0.97	1	0.99	1
	13	0.96	0.34	0.96	1	0.99	1
	11	0.96	0.34	0.97	1	0.99	1
80-20	29	0.96	0.25	0.97	1	0.99	1
	13	0.96	0.38	0.96	1	1	1
	11	0.96	0.38	0.96	1	1	1
10-fold Cross Validation	29	0.96	0.27	0.97	0.99	0.99	0.99
	13	0.96	0.37	0.97	0.99	0.99	0.99
	11	0.96	0.38	0.97	0.99	0.99	0.99

As shown in Table 4, several machine learning classifiers produced variable levels of accuracy. Five classifiers, including LR, SVM, ET, RF, and GDB, performed extraordinarily well with close to 100% accuracy for a certain amount of attributes, namely 29, 13, and 11. RF, ET, and GDB classifiers quantified accuracy with the lowest value using 27 features and highest using 11 features. This produces highly optimistic results that reflect the model's actual predictive performance. The removal of redundant variables improved the classification accuracy of thyroid disease patients, but overall predicted accuracy may have shrunk. In this context, accuracy is not the best metric for assessing predictive performance; instead, specificity, sensitivity, f1-score, and kappa value are considered.

Table.5. Comparison of Sensitivity with 29, 13 and 11 Features

Train – Test Split	Number of Features	LR	NB	SVM	ET	RF	GDB
50-50	29	0.96	0.99	0.96	1	1	1
	13	0.95	0.99	0.96	1	1	1
	11	0.95	0.99	0.96	1	1	1
66-34	29	0.96	0.98	0.97	1	1	1
	13	0.96	0.99	0.97	1	1	1
	11	0.96	0.99	0.97	1	1	1
80-20	29	0.96	1	0.97	1	1	1
	13	0.96	0.98	0.97	1	1	1

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11341

	11	0.96	0.98	0.97	1	1	1
	5						
10-fold Cross Validation	29	0.96	0.27	0.97	1	1	1
	13	0.96	0.37	0.97	1	1	1
	11	0.96	0.38	0.97	1	1	1

Table.6. Comparison of Specificity with 29, 13 and 11 Features

Train – Test Split	Number of Features	LR	NB	SVM	ET	RF	GDB
50-50	29	0.88	0.09	0.93	0.93	0.94	0.98
	13	0.92	0.11	0.92	0.92	0.94	0.95
	11	0.95	0.11	0.92	0.92	0.92	0.95
66-34	29	0.91	0.08	0.93	0.97	0.95	0.98
	13	0.92	0.10	0.93	0.96	0.93	0.96
	11	0.92	0.10	0.93	0.96	0.91	0.96
80-20	29	1	0.09	0.92	1	0.92	1
	13	0.91	0.10	0.94	0.97	0.95	0.97
	11	0.91	0.10	0.94	0.97	0.95	0.97
10-fold Cross Validation	29	0.96	0.98	0.97	1	1	1
	13	0.96	0.99	0.99	1	1	1
	11	0.96	0.99	0.99	1	1	1

Tables 5, 6, 7, and 8 show the performance analysis for specificity, sensitivity, f1-score, and AUC values in 10 fold cross validations respectively. ET, RF, and GDB achieved 100 % sensitivity, specificity, and f1-score with 11 characteristics using 10-fold cross-validation. Except for NB, all classifiers outperformed in terms of sensitivity, specificity, and f1-score for every combination of feature number and training-testing data split.

Table.7. Comparison of f1-score with 29, 13 and 11 Features

Train – Test Split	Number of Features	LR	NB	SVM	ET	RF	GDB
50-50	29	0.97	0.28	0.98	1	1	1
	13	0.97	0.47	0.98	0.99	1	1
	11	0.97	0.47	0.98	0.99	0.99	1
66-34	29	0.98	0.27	0.98	1	1	1
	13	0.98	0.45	0.98	1	1	1
	11	0.98	0.45	0.98	1	0.99	1
80-20	29	0.98	0.32	0.98	1	1	1
	13	0.98	0.51	0.98	1	1	1
	11	0.98	0.50	0.98	1	1	1
10-fold Cross Validation	29	0.96	0.42	0.97	1	1	1
	13	0.96	0.52	0.97	1	1	1
	11	0.96	0.53	0.97	1	1	1

Table.8. Comparison of AUC Value with 29, 13 and 11 Features in 10-Fold Cross Validation

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11341

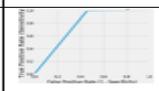
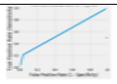
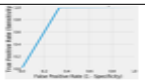
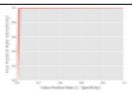
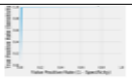
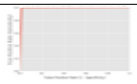
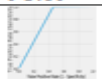
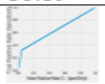
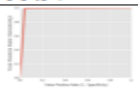
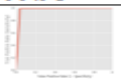
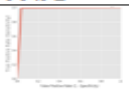
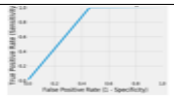
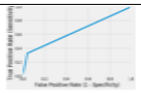
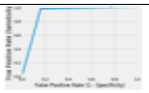
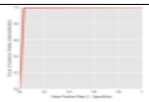
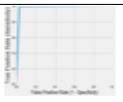
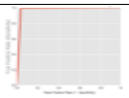
Number of Features	LR	NB	SVM	ET	RF	GDB
All(29)						
	98.69	86.89	99.10	99.97	99.91	99.92
13						
	98.76	90.10	98.78	99.51	99.51	99.37
11						
	98.76	90.20	98.70	99.50	99.51	99.65

Table.9. Comparison of Kappa Statistic with 29, 13 and 11 Features

Train – Test Split	Number of Features	LR	NB	SVM	ET	RF	GDB
50-50	29	0.57	0.03	0.66	0.94	0.94	0.96
	13	0.56	0.06	0.65	0.93	0.95	0.94
	11	0.56	0.06	0.65	0.94	0.92	0.94
66-34	29	0.63	0.02	0.71	0.97	0.95	0.97
	13	0.63	0.05	0.70	0.96	0.94	0.96
	11	0.63	0.05	0.70	0.96	0.94	0.96
80-20	29	0.64	0.03	0.73	0.98	0.95	0.98
	13	0.64	0.05	0.70	0.97	0.96	0.97
	11	0.64	0.05	0.70	0.97	0.96	0.96
10-fold Cross Validation	29	0.66	0.03	0.76	0.96	0.96	0.98
	13	0.66	0.06	0.81	0.95	0.95	0.95
	11	0.66	0.06	0.80	0.95	0.95	0.95

The ROC-AUC score, shown in Table 8, is a composite statistic used to compare the performance of different classifiers and has offered more clarity than accuracy, sensitivity, and specificity. The Kappa statistic calculates the rate of agreement between the expected and predicted outcomes, with values ranging from (1.0), (0.81-0.99), (0.61-0.80), (0.41-0.60), (0.21-0.40), (0.1-0.20) to (0) representing perfect, near-perfect, substantial, moderate, fair, slight, and close to chance agreements. Table 9 shows that all classifiers with 11 features and 10-fold cross validation verified the strong agreement in terms of kappa value. Overall, ET, RF and GDB provided the best accuracy, precision, and specificity with 11 features.

Table.14. Comparison of Performance for Class as Target Variable

Reference	Methods and Accuracy	Sensitivity	Specificity	F1-Measure	Kappa Statistics	AUC
Banu, 2016[33]	LDA: 0.999	0.996	×	×	×	0.997
Tyagi, Mehra & Saxena, 2018[34]	KNN=0.99 SVM==0.99 DT=0.75	×	×	×	×	×
Singh, 2019[35]	SVC: 0.992 NBC: 0.19 KNN: 0.993 MPC: 0.993 DCT: 0.997	×	×	SVC: 0.9957 NBC: 0.28 KNN: 0.996 MPC: 0.996 DCT: 0.998	×	×
Jamkhandikar & Priya, 2020[36]	SVM =0.82 Naïve Bayes=0.83 KNN=0.85	×	×	×	×	×

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11341

Çiçek & Kucukakcali, 2020[37]	0.922	1	×	0.959	×	×
Duggal & Shukla, 2020[38]	NB: 0.74 RF: 0.78 SVM: 0.93	×	×	×	×	×
Shankar et al., 2020[39]	MKSVM: 0.98	0.987	0.96	×	×	×
This Study	**1**	**1**	**1**	**1**	**0.97**	**1**

Table 14 includes a full comparison of the proposed model with past research using Class as the sole target variable. In terms of all performance aspects, the suggested model outperforms the estimates of existing research studies.

CONCLUSION

In terms of accuracy, sensitivity, and specificity, various supervised machine learning algorithms produced varying outcomes. Because of asynchronous values of the test parameters, the relevance of the features varied with the methodologies, causing treatment of major disorders such as TD to suffer from insufficient testing, over-examination, or misunderstanding. This article developed an algorithm for searching for the appropriate attributes while maintaining high levels of accuracy. The dataset with large number of features was mined using GA-based iterations with a probabilistic disease prediction (using Logistic Regression - LR), proposing the set of attributes that improved the accuracy. LR outcomes like any statistical approach are skewed because apparent error rates may underestimate the real value since the model prefers to concentrate on the observed points. As a result, these points may incorrectly depict an optimistic view of the model's genuine accuracy. To address this issue the dataset with reduced number of features is recommended in the first step of iteration was exposed to the second round. This time, the features were decreased by more than half, although the accuracy remained the same. When compared to earlier experiments, the use of proven data mining techniques on the second dataset(with reduced features) resulted in the best degree of accuracy. This research work pioneers a multi-stage, multi-step iterative technique to diagnose TD that may be applied to diagnose any disease. The phases and processes are carefully selected based on literature, constraints, and usefulness. However, because it is an iterative strategy, the time-bound stopping criterion may produce local optima unless the accuracy is near 100%. In such circumstances, the algorithm must be fine-tuned using various train-test splits.

The ability of the proposed ensemble model to select the optimum possible ratio on the training dataset versus the testing dataset is one of its most essential aspects. The suggested model finds the best combination of both sets based on the stated ratio while also experimenting to discover an accurate rule using the ensemble approach. Test findings reveal that the proposed ensemble learning classification model is beneficial in enhancing performance metrics and classification accuracy when compared to its foundation learner and other independent learners stated in the literature. The following are some of the potential consequences of the current research study and the suggested framework.

1. When TD is detected, it is typically followed by a range of normal medical tests performed in laboratories in the presence of specialists or doctors, or during hospitalization. However, this is typically a costly and time-consuming technique. This suggested model incorporates various features taken from daily lifestyles and a few medical test reports from laboratories in text or number format to predict diseases with improved accuracy.
2. This suggested methodology is meant to help medical service providers or doctors give accurate TD diagnoses based on fewer precise and explanatory test results from patients. As a consequence, with the aid of this model, the primary consumer (i.e., medical practitioners or physicians) may forecast TD more quickly and precisely (especially in situations of clinical assumption) and efficiently identify the disease's risk level. This suggested model can function as an electronic doctor, allowing disease to be identified even when medical practitioners are not available. As a result, it has the potential to save lives while also lowering medical expenses substantially.

REFERENCES

1. Thyroid disorders in India: An epidemiological perspective. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3169866/#ref

DOI: 10.17148/IJARCCE.2022.11341

2. Miller, Kimberly D., Rebecca L. Siegel, Chun Chieh Lin, Angela B. Mariotto, Joan L. Kramer, Julia H. Rowland, Kevin D. Stein, Rick Alteri, and Ahmedin Jemal. "Cancer treatment and survivorship statistics, 2016." CA: a cancer journal for clinicians 66, no. 4 (2016): 271-289.
3. Shroff, Shanu, Siddhi Pise, Pratiksha Chalekar, and Suja S. Panicker. "Thyroid disease diagnosis: A survey." In 2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO), pp. 1-6. IEEE, 2015.
4. Thyroid Cancer: https ://seer.cance r.gov/statf acts/html/thyro .html. Accessed 01 Jan 2020
5. Thyroid Problems: https ://medli neplu s.gov/thyro iddiseases.html. Accessed 01 Jan 2020
6. What Is Thyroid Cancer: https ://www.cance r.org/cance r/thyroid-cancer/about /what-is-thyro id-cance r. Accessed 01 Jan 2020
7. Pal, Rekha, Tanvi Anand, and Sanjay Kumar Dubey. "Evaluation and performance analysis of classification techniques for thyroid detection." International Journal of Business Information Systems 28, no. 2 (2018): 163-177.
8. Acharya, U. Rajendra, Pradeep Chowriappa, Hamido Fujita, Shreya Bhat, Sumeet Dua, Joel EW Koh, L. W. J. Eugene, Pailin Kongmebhol, and Kwan Hoong Ng. "Thyroid lesion classification in 242 patient population using Gabor transform features from high resolution ultrasound images." Knowledge-Based Systems 107 (2016): 235-245.
9. Chandel, Khushboo, Veenita Kunwar, Sai Sabitha, Tanupriya Choudhury, and Saurabh Mukherjee. "A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques." CSI transactions on ICT 4, no. 2 (2016): 313-319.
10. Turanoglu-Bekar, Ebru, Gozde Ulutagay, and Suzan Kantarcı-Savas. "Classification of thyroid disease by using data mining models: a comparison of decision tree algorithms." Oxford Journal of Intelligent Decision and Data Sciences 2 (2016): 13-28.
11. Prasad, V., T. Srinivasa Rao, and M. Babu. "Thyroid disease diagnosis via hybrid architecture composing rough data sets theory and machine learning algorithms." Soft Computing 20, no. 3 (2016): 1179-1189.
12. Ray, Arkadip, and Avijit Kumar Chaudhuri. "Smart healthcare disease diagnosis and patient management: Innovation, improvement and skill development." Machine Learning with Applications 3 (2021): 100011.
13. Chaudhuri, Avijit Kumar, Arkadip Ray, Dilip K. Banerjee, and Anirban Das. "A Multi-Stage Approach Combining Feature Selection with Machine Learning Techniques for Higher Prediction Reliability and Accuracy in Cervical Cancer Diagnosis." International Journal Of Computing and Digital System (2021).
14. Steyerberg, E. W. "Validation of prediction models." In Clinical prediction models, pp. 299-311. Springer, New York, NY, 2009.
15. Babyak, Michael A. "What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models." Psychosomatic medicine 66, no. 3 (2004): 411-421.
16. Mostafa, Salama A., Aida Mustapha, Mazin Abed Mohammed, Raed Ibraheem Hamed, N. Arunkumar, Mohd Khanapi Abd Ghani, Mustafa Musa Jaber, and Shihab Hamad Khaleefah. "Examining multiple feature evaluation and classification methods for improving the diagnosis of Parkinson's disease." Cognitive Systems Research 54 (2019): 90-99.
17. Chaudhuri, Avijit Kumar, Dilip K. Banerjee, and Anirban Das. "A Dataset Centric Feature Selection and Stacked Model to Detect Breast Cancer." International Journal of Intelligent Systems and Applications (IJISA) 13, no. 4 (2021): 24-37.
18. Chaudhuri, Avijit Kumar, Deepankar Sinha, Dilip K. Banerjee, and Anirban Das. "A novel enhanced decision tree model for detecting chronic kidney disease." Network Modeling Analysis in Health Informatics and Bioinformatics 10, no. 1 (2021): 1-22.
19. Cavallaro, Gabriele, Morris Riedel, Matthias Richerzhagen, Jón Atli Benediktsson, and Antonio Plaza. "On understanding big data impacts in remotely sensed image classification using support vector machine methods." IEEE journal of selected topics in applied earth observations and remote sensing 8, no. 10 (2015): 4634-4646.
20. Suthaharan, Shan. "Machine learning models and algorithms for big data classification." Integr. Ser. Inf. Syst 36 (2016): 1-12.
21. Breiman, Leo. "Random forests." Machine learning 45, no. 1 (2001): 5-32.
22. Dauwan, Meenakshi, Jessica J. van der Zande, Edwin van Dellen, Iris EC Sommer, Philip Scheltens, Afina W. Lemstra, and Cornelis J. Stam. "Random forest to differentiate dementia with Lewy bodies from Alzheimer's disease." Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 4 (2016): 99-106.
23. Chen, Xi, and Hemant Ishwaran. "Random forests for genomic data analysis." Genomics 99, no. 6 (2012): 323-329.
24. Chaudhuri, Avijit Kumar, Arkadip Ray, Dilip K. Banerjee, and Anirban Das. "An Enhanced Random Forest Model for Detecting Effects on Organs after Recovering from Dengue." methods 8, no. 8 (2021).
25. Verma, Anurag Kumar, Saurabh Pal, and Surjeet Kumar. "Prediction of skin disease using ensemble data mining techniques and feature selection method—a comparative study." Applied biochemistry and biotechnology 190, no. 2 (2020): 341-359.

DOI: 10.17148/IJARCCE.2022.11341

26. Maier, Oskar, Matthias Wilms, Janina von der Gablentz, Ulrike M. Krämer, Thomas F. Münte, and Heinz Handels. "Extra tree forests for sub-acute ischemic stroke lesion segmentation in MR sequences." Journal of neuroscience methods 240 (2015): 89-100.
27. Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." Annals of statistics (2001): 1189-1232.
28. Dodd, Seetal, Michael Berk, Katarina Kelin, Qianyi Zhang, Elias Eriksson, Walter Deberdt, and J. Craig Nelson. "Application of the Gradient Boosted method in randomised clinical trials: Participant variables that contribute to depression treatment efficacy of duloxetine, SSRIs or placebo." Journal of affective disorders 168 (2014): 284-293.
29. Xie, Jianjun, and Stephen Coggeshall. "Prediction of transfers to tertiary care and hospital mortality: A gradient boosting decision tree approach." Statistical Analysis and Data Mining: The ASA Data Science Journal 3, no. 4 (2010): 253-258.
30. Chen, Yifei, Zhenyu Jia, Dan Mercola, and Xiaohui Xie. "A gradient boosting algorithm for survival analysis via direct optimization of concordance index." Computational and mathematical methods in medicine 2013 (2013).
31. Rubini, L. Jerlin, and P. Eswaran. "Generating comparative analysis of early stage prediction of Chronic Kidney Disease." International Journal of Modern Engineering Research (IJMER) 5, no. 7 (2015): 49-55.
32. Vandewiele, Gilles, Isabelle Dehaene, György Kovács, Lucas Sterckx, Olivier Janssens, Femke Ongenae, Femke De Backere et al. "Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling." Artificial Intelligence in Medicine 111 (2021): 101987.
33. Banu, G. Rasitha. "Predicting thyroid disease using linear discriminant analysis (LDA) data mining technique." Commun. Appl. Electron.(CAE) 4 (2016): 4-6.
34. Tyagi, Ankita, Ritika Mehra, and Aditya Saxena. "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, 2018.
35. Singh, Amardip Kumar. "A comparative study on disease classification using machine learning algorithms." In Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE). 2019.
36. Ibrahim, Ibrahim, and Adnan Abdulazeez. "The role of machine learning algorithms for diagnosing diseases." Journal of Applied Science and Technology Trends 2, no. 01 (2021): 10-19.
37. ÇİÇEK, İpek BALIKÇI, and Zeynep KÜÇÜKAKÇALI. "Classification Of Hypothyroid Disease With Extreme Learning Machine Model." The Journal of Cognitive Systems 5, no. 2: 64-68.
38. Duggal, Priyanka, and Shipra Shukla. "Prediction of thyroid disorders using advanced machine learning techniques." In 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 670-675. IEEE, 2020.
39. Shankar, K., S. K. Lakshmanaprabu, Deepak Gupta, Andino Maseleno, and Victor Hugo C. De Albuquerque. "Optimal feature-based multi-kernel SVM approach for thyroid disease classification." The journal of supercomputing 76, no. 2 (2020): 1128-1143.