[1]Amutha C.,
[2]Dr. P. Periyasamy

A Rule-Generative Predictive Clustering Model For Thyroid Prediction Using Learning Approaches

Journal of Electrical Systems

Abstract- Around the world, thyroid disease is considered one of the most highly detected endocrinopathies; for the overall health of a human, thyroid illness is regarded as a significant concern since the thyroid gland controls the human body's metabolism. A dependable, automatic and precise machine-learning (ML) system for thyroid detection is necessary to save time and lower mistake rates. The proposed strategy seeks to overcome previous work's constraints, such as improper comprehensive feature analysis, prediction accuracy improvement, dependability and visualization. Here, 29 clinical factors from a public dataset on thyroid disorders from the California University, Irvine ML repository were employed. By examining symptoms in the early stages and displacing the manual examination of these characteristics, the medical features enabled us to develop a Machine Learning (ML) model that may forecast thyroid-related diseases. Understanding the purpose of features in the prediction tasks of the thyroid is made more accessible with visualization and feature analysis. Furthermore, data balancing and 5-fold CV with using the synthetic minority oversampling technique, intend to solve the over-fitting issue. Because numerous classifiers are involved in the prediction task, learning maintains the trustworthiness of the thyroid forecasting system. Using the proposed Rule-based-Generative Clustering (RGC) with the k-CV method, the suggested model achieved a specificity of 99%, an accuracy of 99.1% and a sensitivity of 99%. This makes it suitable for real-time diagnostic schemes to facilitate disease identification and encourage early-stage treatment.

Keywords- Thyroid prediction, learning, clustering, rule generation, accuracy

1. INTRODUCTION

The substantial rise in human metabolism, reproductive activity and neuronal growth are regulated mostly by thyroid hormones. Regular physical functions: humans are impacted when the thyroid gland cannot produce an optimal hormone level regularly [1]. This is termed thyroid dysfunction. From a medical perspective, thyroid issues can lead to thyroid cancer and thyroiditis. The two chief thyroid-related disorders are hypothyroidism and hyperthyroidism [2]. Globally, the prevalence of thyroid disorders is increasing, leading up to thirty to forty percent of patients assessed in endocrine institutions [3]. Roughly sixty percent of Americans do not know they have a thyroid condition, which affects an estimated 20 million people [4]. Since the difficulties caused by the thyroid illness may be difficult to distinguish from other disorders. It is a challenging process and consumes too much time. The traditional method to diagnose thyroid disorder comprises numerous blood works and proper examination by medical experts. However, the primary concern is accurately diagnosing the illness when it is still in its initial phases [5]. Because blood tests can provide vital information on many thyroid hormones, doctors use factors like triiodothyronine (T3), thyroid-stimulating hormone (TSH), thyroid-stimulating immunoglobulin (TSI) and thyroxin (T4) for suspecting thyroid dysfunction [6]. Twenty-nine clinical features in the dataset were selected for the model's training. Since the data are collected in real-time, pre-processing is necessary to acquire relevant inputs for the ML classifier [7]. Data cleaning, resampling, normalizing and encoding are all part of the data pre-processing step. SMOTE algorithm can be used for data balancing to resolve imbalances in the dataset and reduce bias [8]. By utilizing K-nearest neighbors, SMOTE creates synthetic samples for the minority class. Cross-validation can also be used to check for over-fitting and ensure the models aren’t adequately employed in the training samples and may be extended to newly incoming data [9]. The previous study concentrated on the diagnosis of hypothyroidism among the many thyroid conditions, which include thyroiditis, goiter, thyroid

[1]Research Scholar, Department Of Computer Science, Parks College (Autonomous), Chinnakarai, Tripur, Affiliated to Bharathiar University, Coimbatore, Tamilnadu, India, arjammusan24@gmail.com
[2]Associate Professor, Department Of Computer Science, Parks College (Autonomous), Chinnakarai, Tripur, Affiliated to Bharathiar University, Coimbatore, Tamilnadu, India, drperi1975@gmail.com

malignancy, etc. However, the correct database to train the model and architecture can also be used to address other diseases [10]. The significant contributions of the presented study.

The creation of an automated, straightforward, accurate and ML thyroid prediction framework can be integrated. The proposed Rule-Generative Clustering (RGC) was applied to comprehend the function of different clinical features in thyroid risk prediction.

1. Includes 29 characteristics that, by examining early symptoms, can be used to predict and categorize thyroid conditions. Moreover, it takes the place of laborious manual examination of these characteristics. The SMOTE was used to balance data and guarantee that the outcomes were impartial.
2. The proposed model uses a generative rule concept rather than just one machine learning algorithm; it guarantees the prediction model's reliability.

The work is summarized as follows: section 2 compares various prevailing approaches. The methodology is drafted in section 3, and numerical outcomes are in section 4. The outcomes are summarized in section 5.

2. RELATED WORKS

Conventional machine learning techniques: The researchers developed a thyroid disease prediction framework [11] that employs "multi-kernel SVM," a kernel-based classification technique. Optimizing feature selection was done with a population-based meta-heuristic approach that emulated the gray wolves' hunting strategy and leadership quality and improved gray wolf optimization (GWO) [12] to increase the classification process's efficacy. The population is guided to advantageous areas of the search space by the three aptest candidate solutions: beta, delta and alpha about GWO. We found the ideal feature combination by scanning the feature space adaptively. This method attained 94.5% specificity, 99.05% sensitivity, and 97.49% accuracy [13]. The authors suggest that new and powerful processes are necessary to increase performance and enable the identification of thyroid disorders, as existing techniques necessitate extensive computations. To enhance the disease prediction with the help of the given parameters in the dataset, the authors in [14] used an experiential method to compare the results of DT, RF, ANN, and K-NNs. Additionally, modification of the dataset is done to provide an accurate prediction and it was done on both unsampled and sampled datasets to enhance the comparison outcomes. Once the dataset was modified, the RF algorithm achieved 91% specificity and 94.8% accuracy [15].

Jha et al. [13] employed machine learning (ML) approaches, like RF, LR, DNN, GBM and SVM to identify compounds that have a high probability of initiating thyroid hormone homeostasis. For further testing, the early identification of these chemicals is beneficial in identifying thyroid illness. The ToxCast database offers details on molecular occurrences. With F1-values of 0.81 and 0.83, the TR and TPO showed the most positive predictive outcomes. Three feature-choosing approaches, principal component analysis (PCA), univariate feature selection (UFS) and recursive feature extraction (RFE) were examined in another study [14] about ML algorithms. While UFS chooses the most powerful features and recursive feature extraction eliminates the least robust features till a predetermined feature count is attained, principal component analysis [15] converts data into lower dimensions from high-dimension. Univariate feature selection is an extremely utilized model for certain strategy; RFE is done by constantly selecting the essential features. With an accuracy of 99.35%, the Recursive Feature Elimination and ML classifiers exhibited the best outcomes. Nevertheless, the sample size (n = 519) was limited. A substantial dataset was needed to evaluate the method's effectiveness [16] – [17].

Comprehensive evaluations of prediction thyroid disorders by utilizing several Machine Learning classifiers, both by involving and without involving feature selection strategies, were carried out in another study [18]. It differed from the other studies include extra features like blood pressure, pulse rate, and BMI characteristics. Feature selection in L2 and L1 norms was used in one experiment set, and the other was performed without it [19]. Here, L2 regularization reduced weight magnitudes to deal over-fitting where feature selection is substantially achieved with L1-regularization by eliminating unnecessary features. L1 performs well when a feature subset is significant, while L2 retains a broader collection of traits. With the L2 selection and NB classifier, the final model attained a perfect accuracy of 100%. Another study [20] explored several k-NN classifiers for distance functions in the same database. They also performed a selection of features using the test called chi-squared and L1 norm. Features correlated highly with the target variable are found using the chi-squared test [21]. When handling discrete or category data, this is helpful. Not only does it increase interpretability and avoid over-fitting, but it also improves

the model's performance. 100% accuracy was attained using chi-squared methods with KNN, cosine distance functions and Euclidean. Nevertheless, a public thyroid database has not yet been used to validate the suggested model. Therefore, its performance in real-time scenarios and over-fitting areas is not verified.

Santos et al. [22] utilized ML methods, like RF, DT, K-star classifiers and sequential minimum optimization (SMO), to predict hypothyroidism-related diseases. During SVM training, large-scale quadratic programming optimization problems are solved. SMO breaks these lengthy optimization processes into more manageable, analytically targeted tasks [23]. The proposed work handled three thousand seven hundred seventy-two different records. The outcomes are achieved with RF and DT with 98 % and 99% accuracy respectively. However, the current study did not take hyperthyroid prediction into account. There has been a lot of discussion on how well-supervised and unsupervised classifiers predict thyroid problems. Different research [24] focused on the concept of features that are relevant and medical applications, deviating from this pattern. They enumerated the four features considered the best and most helpful in determining thyroid disorder and showed how simple and inexpensive it was for practitioners to assess these features. Additionally, they outlined the disadvantages of not attempting to test the entire thyroid panel, which is a common practice in various countries. In conclusion, the outcomes exhibit stability and are unlikely to change based on the classifier employed or the underlying features of the dataset, including imbalance.

Prediction accuracy has to be improved in the existing studies investigated based on traditional methodologies because prediction accuracy measures are essential when considering machine learning in the medical field [25]. Furthermore, there was no evidence for reassessing the models' reliability because they were based only on one machine learning classifier. CAD systems can only use the suggested models once they are proven reliable. While some studies have demonstrated excellent accuracy, their development was based on private rather than public datasets, and their data acquisition was subject to permissions, privacy, secrecy, and other issues [26]. With DL networks, the investigation is done with wider training data, powerful computations, black-box nature and longer training periods. Moreover, the essential characteristics of decision-making still need to be discovered. This work proposes a system with adequate trustworthiness by utilizing a classifiers ensemble, in which many classifiers contribute to the decision-making of diagnosis to address these problems. The classic ML algorithms guaranteed the model's ease of use and performance. Furthermore, the attributes are accessible and examined at any point. By using public data to train the model, it will be able to handle a broader range of data and be integrated into CAD systems in real-time scenarios [27] – [28].

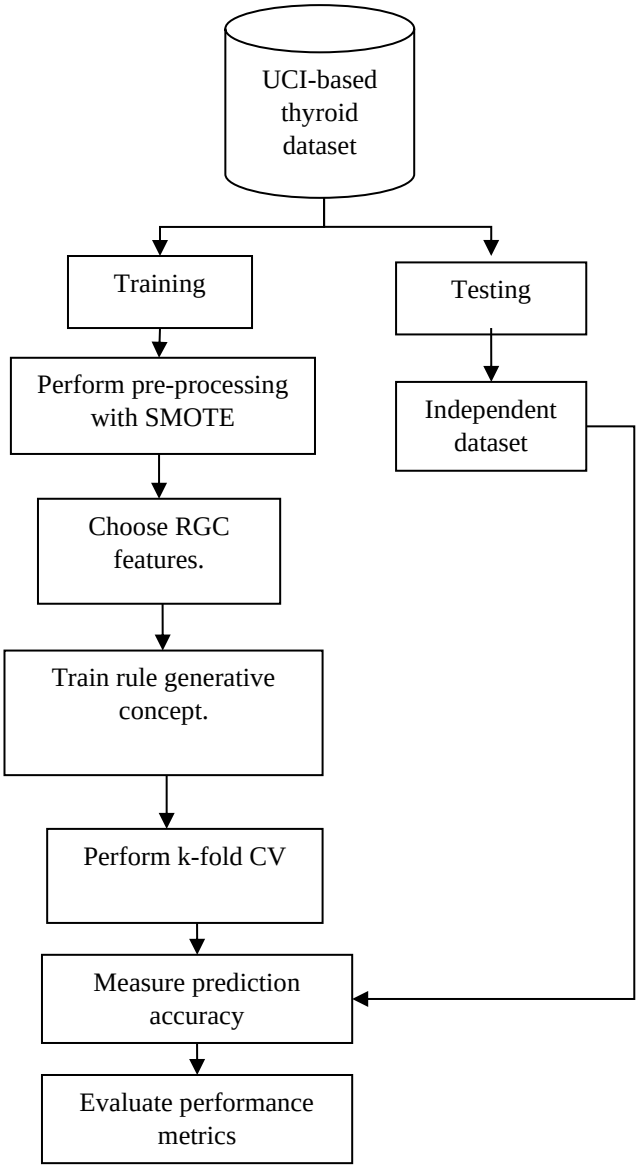
Table 1: Comparison of existing approaches

Ref	Method	Input variable	Output	Performance evaluation
[16]	Customized Alexnet	29 clinic pathological characteristics	Thyroid peroxidase (TPO) active, inactive	91% specificity and 94.7 % Accuracy.
[18]	ML + RFE classifier	Molecular descriptors	Hypothyroid, non-hypothyroid	F1-score 0.83
	L2 selection + NB classifier	Age, ID, sex, FT4, FT3, T4, TSH and T3	Hypo- and hyper-thyroid, and normal	99.45% Accuracy
[15]	multi-kernel + SVM Gray wolf optimization	clinicopathological characteristics-29	Hyperthyroid, normal and hypothyroid	97.49% Accuracy, 94.5% specificity and 99.05% sensitivity
[14]	SMO + RF	Gender, ID, BMI, age, pulse rate, pregnancy, BP, TSH, T4 and T3	Hypothyroid, non-hypothyroid	100% Accuracy
[19]	CNN	29 clinicopathological characteristics	Thyroiditis, cystic, normal, adenoma, cancer and multi-nodular goiter	100% Accuracy
[22]	KNN+ Chi-square test	Gender, ID, BMI, age, pulse rate, pregnancy, BP, TSH, T4 and T3	Hyperthyroid, normal and hypothyroid	100% Accuracy

3. METHODS AND MATERIALS

3.1. Database

The utilized thyroid dataset was accessed from the UCI ML repository [29]. Six databases totaling 2800 training and 972 test examples are included, all from the Garvan Institute, Australia. This was applied to categorize binary data. The predictive variable in the database was labeled as categorical, with "P" denoting hypothyroidism and "N" indicating the opposite. The dataset contains several features corresponding to the patients' various clinicopathological traits. There was a total of 29 actual or category attributes present. The characteristics include sex, age, antithyroid drugs,thyroid_surgery, query_hyperthyroid, I131_treatment, pregnancy, query_hyperthyroid, query_hypothyroid, query on thyroxine and query_hypothyroid. The following blood parameters were taken into consideration: TSH (the genuine TSH value), TSH_measured, TT4 referral_source, T3, T4U, TT4_measured, TBG, T3_measured, FTI_measured, FTI, lithium, T4U_measured, TBG_measured, tumor, hypo pituitary, and goiter. The target variable employed in this study was hypothyroidism.



The flow of the classifier model

3.2. SMOTE

Issues with data imbalance occur in datasets in real-time scenarios. The SMOTE was employed to deal with thyroid database's data imbalance problem. The minority class oversampling is a popular solution to the imbalanced dataset issue. However, duplicating current instances from a minority class might not offer fresh perspectives. Instead, SMOTE uses the knowledge from preexisting examples to create new synthetic ones. In the feature space, samples near one another are selected via the SMOTE approach. It makes a line that connects the sample chosen to its five closest neighbors, on average, and afterward creates fresh samples. By bridging the gap between two examples and their arbitrarily chosen neighbors, this synthetic example efficiently raises the minority class's depiction. We reduced the issue of over-fitting that can result from random oversampling by using SMOTE. By ensuring that synthetic examples are produced systematically relies on the minority class instances distribution that already exist, this strategy enhances the model's capacity for generalization.

3.3. Prediction with Rule-Generative Clustering (RGC)

The K-means clustering technique pre-processes the data and clusters the source samples after collecting them. Additionally, a rule for data encoding is suggested for the clustered data sample's encoding and streamlining of the data form. This rule works with K-means clustering to perform data mining in the first layer and sort the samples to create a sample library. The following are the guidelines and procedures. The proposed clustering technique determines three factors: data samples are classified with Euclidean distance; the clustering is assessed with criterion function; and optimal classification clusters are found with coefficients evaluation. To achieve sample classification, the clustering approach adopts Euclidean distance among each sample data and cluster center. The samples are then placed into the cluster with the smallest Euclidean distance. Eq. (1) can be used to compute the Euclidean distance.

$$d(X,Y^i) = \sqrt{\left(x_1 - y_1^j\right)^2 + \left(x_2 - y_2^j\right)^2 + \cdots + \left(x_n - y_n^j\right)^2} = \sqrt{\sum_{i=1}^{n}\left(x_i - y_i^j\right)^2} \quad (1)$$

Here, $X = (x_1, x_2, \ldots, x_n)$ is a sample that is unclassified in $n-$dimensional space that associates with the elements (the non-faulty node's voltage data) in the source sample library's. $Y^j = (y_1^j, y_2^j, \ldots, y_n^j)$ is the j^{th} cluster's center. During the first-time sample classification process, whichever sample may be chosen randomly as the cluster center. The cluster center is updated using the mean of all samples within each cluster, and its inconsistency, like not updating, is monitored by the criteria function. Eq. (2) given below shows the criterion function, which aims to minimize the sum of the squared errors between the cluster center and the cluster samples:

$$min\sum_{j=1}^{K}\sum_{x_i^j \in X^j, y_i^j \in Y^j}\left(x_i^j - y_i^j\right)^2 \quad (2)$$

Here, the number of the clusters is K, the j^{th} cluster center is Y_j, the i^{th} element data is y_i^j, x_i^j is the i^{th} element data in X_j, and the j^{th} cluster's random sample is X_j. The cluster center ceases to update when the criteria function of formula Eq. (2) the sample into $K-$clusters is over. The contour coefficients of various clusters are computed and used to find the ideal cluster count for K-means clustering in first-layer data mining. The cluster count with the most significant contour coefficient is then fixed to be the perfect cluster count. The contour coefficient calculation procedure for each sample of a cluster is presented in Eq. (3):

(1) Firstly, the computation of α_k shows whether the cluster cohesion is done. (The average distance between x and every other point in its cluster).

(2) Next, the separation degree b_k is computed among the clusters and the others. (The average distance between all points in different clusters and x).

(3) Finally, the computation of S_k, the contour coefficient, is done. (The difference between b_k and α_k is divided by the most significant value among the two).

$$S_k = \frac{b_k - \propto_k}{\max(b_k, \propto_k)} \tag{3}$$

The contour coefficient's value is in the $[-1, 1]$ range. The S_k value will be more significant if the contour coefficient's value is closer to 1. The existing cluster number $K's$ contour coefficient is determined by averaging the all-sample's contour coefficients. The classification operation is better with a more significant contour coefficient and greater distance between the clusters. Consequently, the ideal cluster number for the source sample library is determined by looking at the K value with the most significant contour coefficient.

3.4. Rule generation

Fig 2 shows the guidelines for applying the K −means clustering approach to clustering in a source sample library: By increasing K from 5, the enumeration approach yields the ideal number of clusters. The following is a description of the clustering rules when there are K clusters:

1) First, K samples are selected randomly to create the initial cluster center.

2) Every sample is sorted into the cluster with the smallest Euclidean distance based on the distance it takes to get to the center of each cluster, as determined by Eq. (1).

3) To establish whether the minimum is obtained, the criterion function is computed by averaging all samples within every cluster, and it is designated as the cluster center.

4). Till the Eq. (2) criterion function approaches the lowest, this process will be repeated.

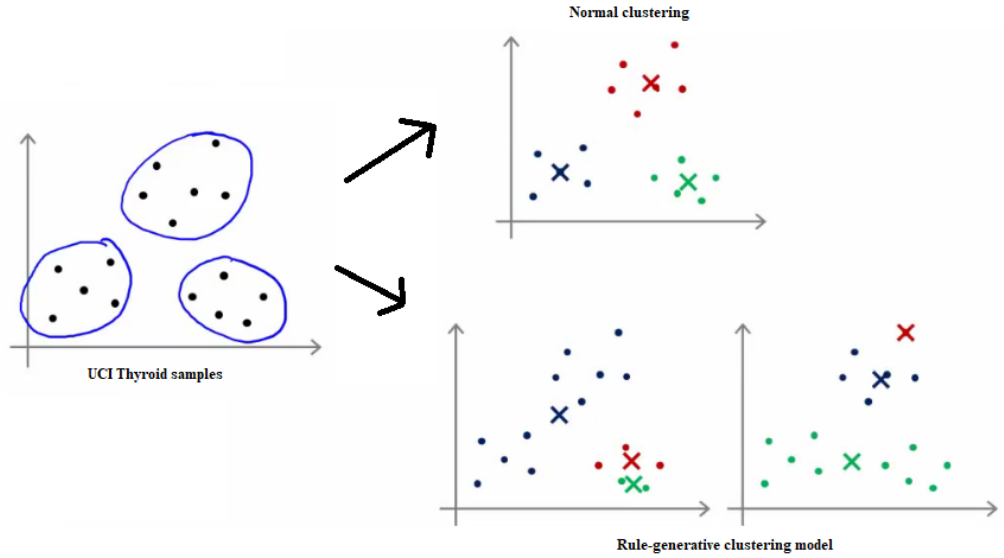


Fig 2: Proposed Rule-generative clustering model

The optimum K values of the clusters on the various nodes differ whenever the clustering of source samples happens based on different failure types. To find its optimal K and finish clustering the source samples on each node, each source sample in a different node must go through the abovementioned procedure. Following the source samples' clustering, there are some similarities among the samples in each cluster. Even though there are some similarities between each cluster's samples, once clustering is done, the data form becomes easier to manage. To simplify the data, self-encoding is applied to the categorized samples following the clustering of the source sample libraries. The following rules are created to preserve the encoded data's critical attribute data, like the cluster and node to which the sample belongs: After clustering, each sample is queried first for the node (T), then for the cluster (W), and $T0W$ is the last coding form. For instance, $T \in N$ (N is an integer that indicates the node

number excluding the defective node), $0 < W \leq K$, and $W \in N$. The $T0W$ value indicates that the sample is the voltage data sample classified into the 1^{st} node 3^{rd} cluster. Coding rules are illustrated in Fig 2 after self-encoding and clustering. Following the source samples' self-encoding and clustering, the sample data take on a more manageable and compact form. The 1st-layer data mining is finished when the source sample library is changed into the sample library by processing the source samples using self-encoding and K-means clustering. It extracts the inner relationships between the source samples and various unlabeled data samples, maximizes the relevance of the data sample within its cluster, and sets up the data mining process for the second layer.

4. EXPERIMENTAL RESULTS AND ANALYSIS

The findings from every experiment while developing the system are discussed and shown. Numerous pre-processing techniques were applied to enhance the system's performance, such as encoding categorical data and identifying and managing missing values. Different machine learning classifier algorithms were tested in addition to hyper-parameter adjustment and cross-validation. The prepared model's effectiveness was assessed by evaluating its accuracy, sensitivity, specificity, and other attributes. Step one started by using a thyroid dataset as the work wanted a database in a real-world scenario with sufficient training data. In routine medical operations, a system created from a real-time database can be helpful. More data is necessary for increased accuracy. Researchers must additionally deal with extra difficulties when utilizing private healthcare data for research, such as getting consent to collect the data, privacy and confidentiality issues, etc. Numerous scholars use this publicly available thyroid disorder database. The SMOTE technique was applied during pre-processing because the dataset had data imbalance issues. A data-mining technique called data pre-processing is employed to make the raw data acquired more useful. The pre-processing is done using the Pandas Python library. Value-containing attributes, including letters and digits, were present in the actual data. To be readable by machines, it needs to be encoded. As a result, data were transformed into data frames, and the Sklearn label encoder was used to encode the data. The "int64" data type was then applied to all 29 attributes. Subsequently, redundant rows were removed from the encoded data to perform data cleaning and guarantee that the input for the classifier would not be redundant. Next, as indicated in Table 2, the summary of statistical data was assessed through standard deviation and mean etc. The term "percentile" refers to the number of values that fall below the given percentile.

Table 2: Data summary

Index	Age	Sex	TSH	T3	TT4	T4U	FTI1
Count	3712	3712	3712	3712	3712	3712	3712
Mean	46	1.2	123	31	120	65	108
SD	20	0.5	82	20	97	32	97
Min	0	0	0	0	0	0	0
25%	27	1	64	17	21	47	18
50%	49	1	112	24	80	58	57
75%	62	2	166	31	225	70	220
Max	92	2	286	70	24	145	235

Table 3: Performance evaluation

Classifier	Class	PPV	NPV	TPR	TNR	Accuracy	Misclassification rate
LR	0	86	88	89	86	87	0.12
	1	87	86	86	89		
NN	0	95	95	95	94	94	0.05
	1	94	94	94	95		
DT	0	96	96	96	94	95	0.04
	1	95	95	94	96		
k-NN	0	94	94	95	88	91	0.08
	1	89	89	88	95		
SVM	0	90	90	91	84	88	0.11
	1	85	85	84	95		
Bagging	0	89	89	91	91	89	0.018

	1	79	79	93	93		
Boosting	0	95	95	95	93	98	0.005
	1	93	93	93	95		
Proposed	0	99	99.1	99	99.5	99	0.004
	1	98.8	98	98.5	99.6		

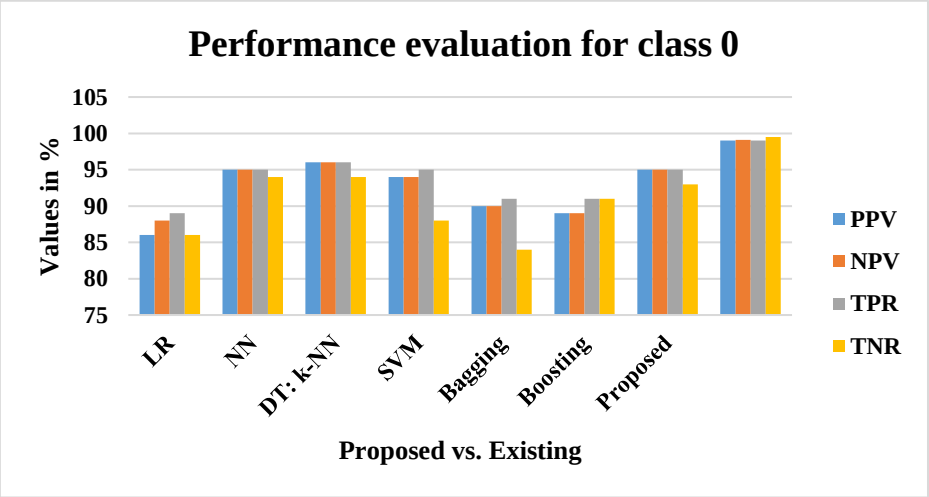


Fig 3: Performance evaluation for class 0

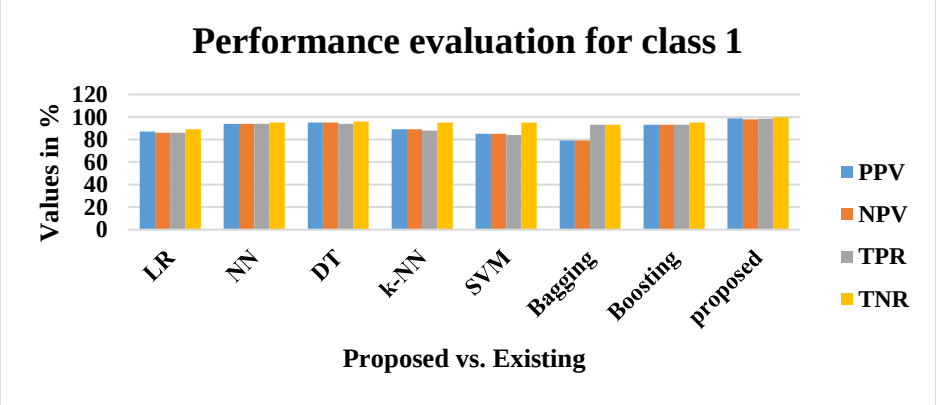


Fig 4: Performance evaluation for class 1

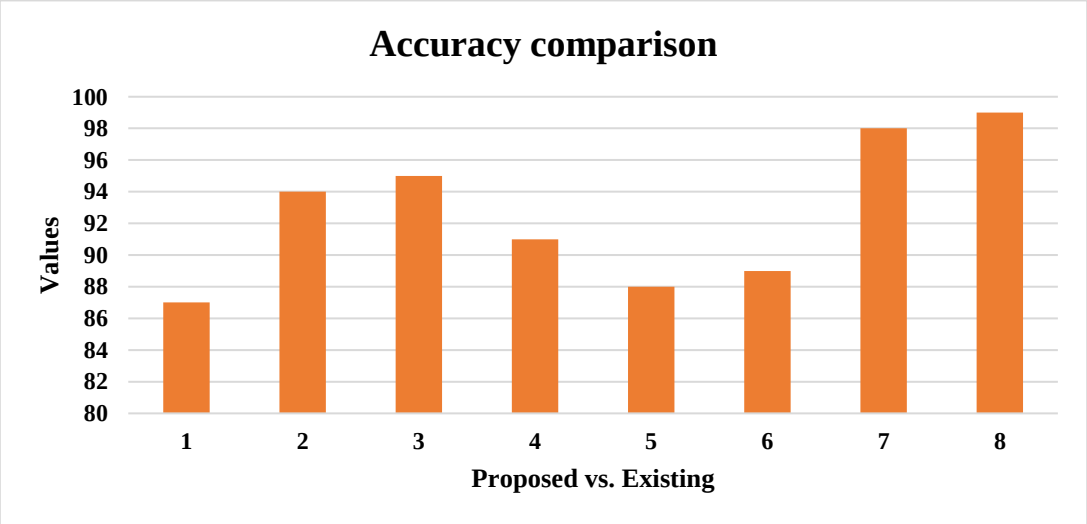


Fig 5: Accuracy comparison

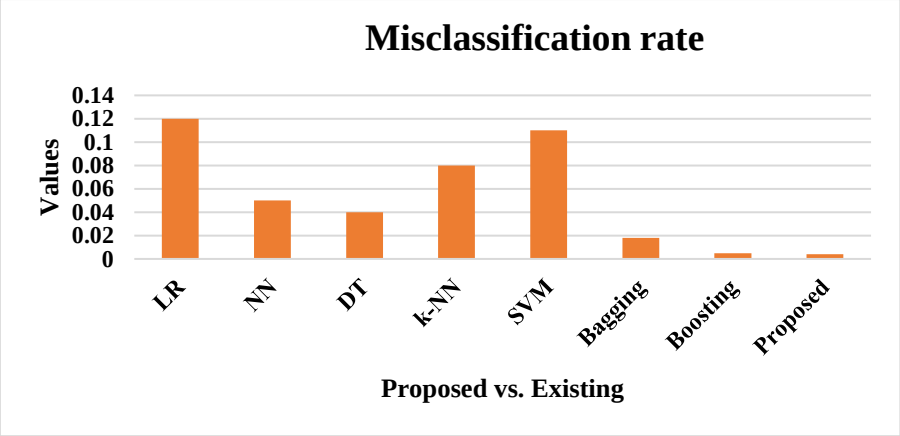


Fig 6: Misclassification rate

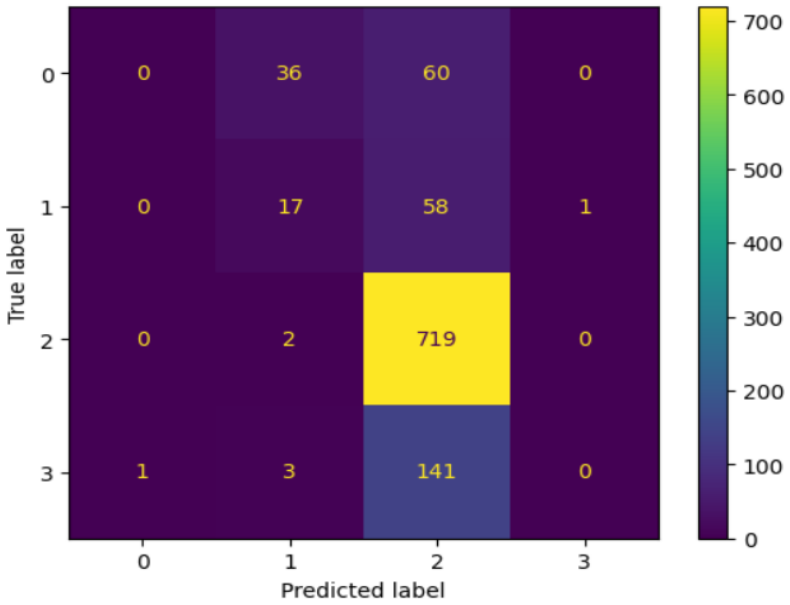


Fig 7: Confusion matrix

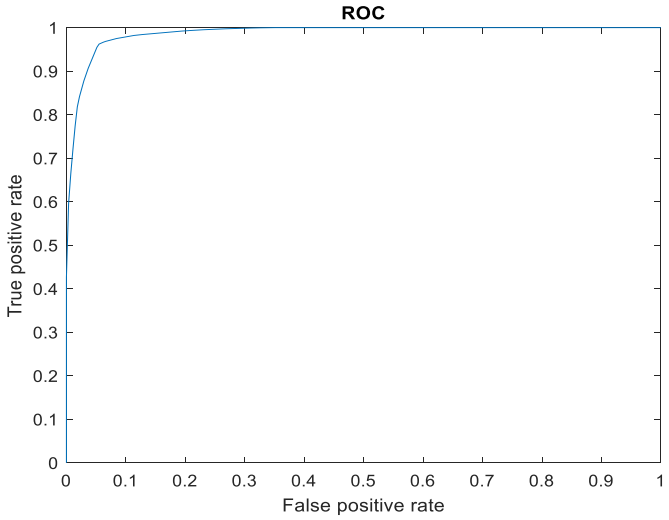


Fig 8: ROC computation

After pre-processing, the target and data features are extracted from the pre-processed data frame. The feature-target data were obtained to use in the ML phases. Using 80:20 split, dataset was separated into eighty percent of data into training sets and the remaining twenty percent into testing sets. After splitting the data using SVM, NN, LR, DT and KNN algorithms with 5-fold cross-validation, we utilized numerous ML models. The DT classifier produced the highest classification accuracy of 95.62%. As the methodology is intended for clinical disease identification, where correctness is crucial, more improvement in prediction accuracy was necessary. Reliability was yet another issue in integrating the model into real-time systems. The RGC learning algorithms were used to meet these goals. They outperformed traditional classifiers, with 99.5% accuracy being the highest level attained through RGC. In this case, non-thyroid classes are indicated by class 0, and thyroid classes are marked by class 1. As shown in Table 2, we assessed the model's true-negative rate (TNR), misclassification rate, PPV, NPV and TPR. Out of all the data samples, accuracy is the number of samples from the test dataset correctly classified. TPR or sensitivity is expressed as the ratio of correctly identified positive cases amongst all positive cases. Eq. (4) to Eq. (8) defines specificity, also known as TNR, as the ability of the model to correctly classify a data sample, including a negative instance amongst the entire negative cases. Precision, also known as PPV, evaluates the performance by correctly predicting the positive cases to those that the methodology predicts. Eq. (4) states that the chance of data sample having a negative screening test that doesn't have a particular disorder is known as the NPV. TN and TP accurately predict positive and negative situations. The terms "false positive" and "false negative" refer to incorrectly identifying positive cases as negative and negative cases as positive.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$TNR = \frac{TN}{TN + FN} \tag{7}$$

$$NPV = \frac{TN}{TN + FN} \tag{8}$$

The models undergo fine-tuning using rule-based clustering and Fig 8 displays the classification error curve. With 7000 obs/s prediction speed, the finalized RGC approach required 148.82 s for training. We exported the FNR and TPR plots for the finished thyroid prediction model by using the RGC scheme to observe the performance of each class's classifier. In this case, FPs are 14, false negatives (FNs) are 3406, true positives (TPs) are 21 and true negatives (TNs) are 3399. The TPR calculates the number of observations correctly classified for every true class. The FNR provides the proportion of incorrectly identified observations to correctly identified observations. There are summaries for each true class in the plot's final two columns to the right. FDR and PPV plots are presented as shown in Fig 4 to Fig 8, as FP values are essential to our classification problem. FDR measures the ratio of wrongly categorized observations for every predicted class, and PPV represents the fraction of correctly classified data in every expected class.

Table 4: Benchmark method comparison

Ref	Method	Result
[16]	**ANN**	Acc = 95%, pre = 95%, recall = 95% and F1-score = 95%
[18]	**RF**	Acc = 94%, pre = 94.5%, recall = 91%
[19]	**MLP**	Acc = 89%
[23]	**NN**	Acc = 98%
[25]	**RF**	Acc = 91%
[26]	**XG-boost**	Acc = 95%, PPV = 95%, NPV = 93%
[28]	**SMOTE**	Acc = 93%, pre = 95% and AUC = 93%
	Proposed	Acc = 99%, pre = 98.9%, recall = 99.1% and F1-score = 99%

Based on the currently selected classifier, the TPR and FPR for different classification score boundaries are shown on the ROC curve. The area under the ROC curve (AUC) value is equivalent to the combination of an ROC curve, considering that FPR is from FPR = zero to FPR = one. The Area under the Curve is used to quantify the efficiency of the classifier as a whole; a value of one, which is the maximum, implies a better performance by the classifier. The perfect separation between negative and positive classes in a model's predicted probability is shown by an AUC of 1.000. However, because different matrices consider variables like trade-offs, class imbalance, and classification threshold, this does not always translate into 100% performance in other metrics. Each performance metric provides different performance features and system behavior. Table 5 compares the suggested model's performance with a few related works experimented on the same database.

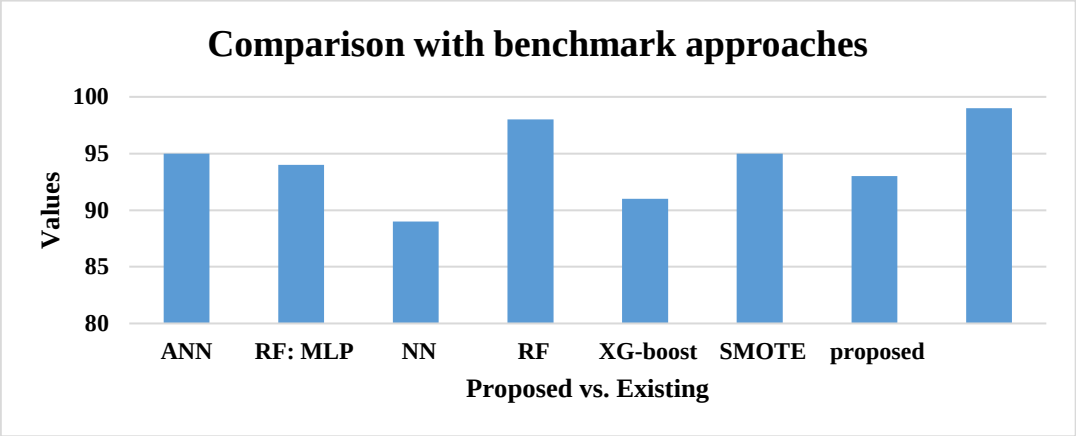


Fig 9: Accuracy comparison with benchmark approaches

Individuals prescribed to the thyroid centers for assessment are in the user database. The user dataset is a valuable tool for assessing and training the proposed methodology since it contains significant medical variables for diagnosing disorders related to the thyroid, which is the primary focus of the prediction model. This would guarantee the model's suitability for this particular task. Techniques like SMOTE and cross-validation can reduce bias and increase generalizability. By retraining the model with a suitable database, this research can also be expanded to many similar clinical disease identification applications. The suggested methodology is helpful for disease identification in real-time because it exhibits numerous essential features and attributes, like improved robustness, accuracy, specificity, sensitivity, adaptability, and the potential to be incorporated into prediction systems. Using the suggested ML model, our approach can be integrated to enable patient data entry and predict the patient's thyroid condition. After a labeled public database becomes available, we will evaluate a multi-class thyroid classification model. Then, RGC is tested in addition to the display of feature relevance in classification decisions.

5. CONCLUSION

Thyroid diseases are increasingly becoming highly prevalent; the discovery of them in their early stages is vital; treatment can prevent complications, and death rates can be decreased; the identification of these disorders has grown in importance within the medical community. Precisely forecasting the disease course and realizing the interaction of medical signs are crucial for clinical disease identification and treatment. Thus, the need for an efficient real-time system arises. Most previous research has concentrated on a single model that does not guarantee increased accuracy or model reliability. In contrast, rule-based essential requirements for ML models are utilized in medical systems. No one knows the primary standards by which the decisions are determined, even though some ML techniques provide more accuracy than the standard ML limitations. Therefore, all of the restrictions are removed in this work by creating a traditional ML model using a Rule-Generative Clustering (RGC) learning strategy and appropriate clinical feature analysis. Foreseeing the disease course and the clinical symptom's interdependence or characteristics is essential to clinical diagnosis and treatment. Together with 99.1% specificity and PPV, 99% sensitivity and NPV, and 100 AUC, the model yielded 99.5% accuracy. The use of approaches in producing binary classification judgments ensures model reliability. Furthermore, building the model using open databases contributes to the model's stability when considering real-time system use.

REFERENCES

[1] Rashad, N.M.; Samir, G.M. Prevalence, risks, and comorbidity of thyroid dysfunction: A cross-sectional epidemiological study. Egypt. J. Intern. Med. 2020, 31, 635–641.

[2] Mirbabaie, M.; Stieglitz, S.; Frick, N.R.J. Artificial intelligence in disease diagnostics: A critical review and classification on the current state of research guiding future direction. Health Technol. 2021, 11, 693–731.

[3] Holzinger, A.; Keiblinger, K.; Holub, P.; Zatloukal, K.; Müller, H. AI for life: Trends in artificial intelligence for biotechnology. New Biotechnol. 2023, 74, 16–24.

[4] Alyas, T.; Hamid, M.; Alissa, K.; Faiz, T.; Tabassum, N.; Ahmad, A. Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach. Biomed. Res. Int. 2022, 2022, 9809932.

[5] Garcia de Lomana, M.; Weber, A.G.; Birk, B.; Landsiedel, R.; Achenbach, J.; Schleifer, K.J.; Mathea, M.; Kirchmair, J. In Silico Models to Predict the Perturbation of Molecular Initiating Events Related to Thyroid Hormone Homeostasis. Chem. Res. Toxicol. 2021, 34, 396–411.

[6] Riajuliislam, M.; Rahim, K.Z.; Mahmud, A. Prediction of Thyroid Disease (Hypothyroid) in Early Stage Using Feature Selection and Classification Techniques. In Proceedings of the 2021 International Conference on Information and Communication Technology for Sustainable Development, ICICT4SD 2021—Proceedings, Dhaka, Bangladesh, 27–28 February 2021; pp. 60–64.

[7] Omuya, E.O.; Okeyo, G.O.; Kimwele, M.W. Feature Selection for Classification using Principal Component Analysis and Information Gain. Expert Syst. Appl. 2021, 174, 114765.

[8] Rehman, H.A.U.; Lin, C.Y.; Mushtaq, Z.; Su, S.F. Performance Analysis of Machine Learning Algorithms for Thyroid Disease. Arab. J. Sci. Eng. 2021, 46, 9437–9449.

[9] Demir-Kavuk, O.; Kamada, M.; Akutsu, T.; Knapp, E.W. Prediction using step-wise L1, L2 regularization and feature selection for small data sets with many features. BMC Bioinform. 2011, 12, 412

[10] Zhai, Y.; Song, W.; Liu, X.; Liu, L.; Zhao, X. A Chi-Square Statistics Based Feature Selection Method in Text Classification. In Proceedings of the IEEE International Conference on Software Engineering and Service Sciences (ICSESS), Beijing, China, 23–25 November 2018; pp. 160–163.

[11] Mishra, S.; Tadesse, Y.; Dash, A.; Jena, L.; Ranjan, P. Thyroid disorder analysis using random forest classifier. Smart Innov. Syst. Technol. 2021, 153, 385–390.

[12] Zhang, X.; Lee, V.C.; Rong, J.; Lee, J.C.; Liu, F. Deep convolutional neural networks in thyroid disease detection: A multiclassification comparison by ultrasonography and computed tomography. Comput. Methods Programs Biomed. 2022, 220, 106823

[13] Jha, R.; Bhattacharjee, V.; Mustafi, A. Increasing the Prediction Accuracy for Thyroid Disease: A Step Towards Better Health for Society. Wirel. Pers. Commun. 2022, 122, 1921–1938.

[14] Prathibha, S.; Dahiya, D.; Robin, C.R.; Nishkala, C.V.; Swedha, S. A Novel Technique for Detecting Various Thyroid Diseases Using Deep Learning. Intell. Autom. Soft Comput. 2023, 35, 199–214.

[15] Alshayeji, M.H.; Sindhu, S.C.; Abed, S. CAD systems for COVID-19 diagnosis and disease stage classification by segmentation of infected regions from CT images. BMC Bioinform. 2022, 23, 264.

[16] Alnaggar, M.; Handosa, M.; Medhat, T.; Rashad, M.Z.; Author, C.; Alnaggar, M. Thyroid Disease Multi-classMulti-class Classification based on Optimized Gradient Boosting Model. Egypt. J. Artif. Intell. 2023, 2, 1–14.

[17] Sengupta, D.; Mondal, S.; Raj, A.; Anand, A. Binary Classification of Thyroid Using Comprehensive Set of Machine Learning Algorithms. In Frontiers of ICT in Healthcare: Proceedings of EAIT 2022; Springer Nature: Singapore, 2023; pp. 265–276.

[18] Trivedi, N.K.; Tiwari, R.G.; Agarwal, A.K.; Gautam, V. A Detailed Investigation and Analysis of Using Machine Learning Techniques for Thyroid Diagnosis. In Proceedings of the 2023 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 1–3 March 2023; pp. 1–5.

[19] Islam, S.S.; Haque, M.S.; Miah, M.S.U.; Sarwar, T.B.; Nugraha, R. Application of machine learning algorithms to predict the thyroid disease risk: An experimental comparative study. PeerJ Comput. Sci. 2022, 8, e898.

[20] Kumari P, Kaur B. Mechanism for disease classification in predicting thyroid disease. In: 2023 1st International conference on circuits, power and intelligent systems (CCPIS); 2023, pp. 1–6.

[21] Zimek A, Filzmoser P. There and back again: Outlier detection between statistical reasoning and data mining algorithms. Wiley Interdiscip Rev Data Min Knowl Discov. 2018;8(6):1–26.

[22] Santos-Pereira J, Gruenwald L, Bernardino J. Top data mining tools for the healthcare industry. J King Saud Univ Comput Inf Sci. 2022;34(8):4968–82

[23] Begum A, Parkavi A. Prediction of thyroid disease using data mining techniques in 2019 5th International conference on advanced computing and communication systems ICACCS 2019; 2019 pp. 342–345.

[24] Zhu Y, Fu Z, Fei J. An image augmentation method using the convolutional network for thyroid nodule classification by transfer learning. In: 2017 3rd IEEE International Conference on Computer Communications ICCC 2017; 2017, vol. 2018-Janua, pp. 1819–1823.

[25] Ilyas M, et al. Deep learning based classification of thyroid cancer using different medical imaging modalities: a systematic review. VFAST Trans Softw Eng. 2021

[26] Begum A, Parkavi A. Prediction of thyroid disease using data mining techniques. In: 2019 5th International Conference on Advanced Computing and Communication Systems ICACCS 2019, no. August 2016; 2019, pp. 342–345.

[27] Ahmed I, et al. Lithium from breast milk inhibits thyroid iodine uptake and hormone production, which are remedied by maternal iodine supplementation. Bipolar Disord. 2021;23(6):615–25.

[28] Priyadharsini D, Sasikala S. Efcient thyroid disease prediction using features selection and meta-classifiers. In: Proceedings—6th International Conference on Computing Methodologies and Communication ICCMC 2022, no. ICCMC; 2022, pp. 1236–1243

[29] Ahmed I, Mohiuddin R, Muqeet MA, Kumar JA, Thaniserikaran A. Thyroid cancer detection using deep neural network. In: Proceedings— International Conference on Applied Artificial Intelligence and Computing, ICAAIC 2022, no. ICAAIC; 2022, pp. 166–169

[30] Jamil M, Wlvvxh WKU. Ensemble-based effective diagnosis of thyroid disorder with various feature selection techniques; 2022, pp. 14–19.