International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(1s) 2026 pp. 363-376
https://doi.org/10.70917/ijcisim-2026-2288
Article

Intelligent Diagnosis of Hypothyroidism Using Hybrid Machine Learning with Optimized Feature Engineering and Parameter Tuning

Swathi Lenka[1], Nirmal Keshari Swain[2], Jui Pattnayak[3], Manish Kumar[4], Davinder Paul Singh[5], Debabala Swain[6], Debabrata Swain[7]

[1] Department of Information Technology, GMR Institute of Technology (Deemed to be University), Rajam, Andhra Pradesh, India. swathi.l@gmritdu.edu.in
[2] Department of Information Technology, Vardhaman College of Engineering, Telangana, India. swain.nirmal6@gmail.com
ORCID: 0000-0002-6845-5808
[3] Department of Information Technology, JIS College of Engineering, Kolkata, West Bengal, India. jui.pattanayak@jiscollege.ac.in
[4] Department of Electronics and Communication Engineering, Pandit Deendayal Energy University, Gujarat, India. Manishkumar087@gmail.com
[5] Department of Computer Science and Engineering, Pandit Deendayal Energy University, Gujarat, India. Davinder.Singh@sot.pdpu.ac.in
[6] Department of Computer Science, Ramadevi Women's University, Odisha, India. debabala.swain@gmail.com
[7] Department of Computer Science and Engineering, Pandit Deendayal Energy University, Gujarat, India. debabrata.swain7@yahoo.com
Corresponding Author: Debabrata Swain, debabrata.swain7@yahoo.com

Abstract: The thyroid gland is considered one of the vital organs in the human body. It secretes hormones crucial for maintaining metabolism. This disease creates several abnormalities in the human body. It has a myopic characteristic that has been unnoticeable for detection. For the detection of the thyroid, several complicated clinical tests are being performed. Sometimes, due to statistical human errors, it is not possible to detect the precise status of the disease. This work, a hybrid machine learning based diagnostic screening system has been proposed for the accurate identification of thyroid disease. For boosting the performance of the system, less correlated features are removed by using the Chi-Square test. The hybrid machine learning-based classifier is formed by stacking different basic algorithms such as Random Forests, SVM, and KNN. The resultant system is optimized by performing hyper-parameter tuning using Grid Search CV. The highest accuracy obtained by the system is 99.79%.
Keywords: Thyroid; Feature selection; Chi-square; Data modelling; Machine learning; Stacking classifier; Grid search CV

1. Introduction

Thyroid has appeared to be one of the globally common ailments in the recent times. One of the most prevalent conditions globally is hypothyroidism, which arises mostly due to iodine deficiency. The Health Examination Survey data for the America showed that the total rate of hypothyroidism is 4.6%. In the United States, screening research found that the prevalence of subclinical hypothyroidism was 9%, and among women 75 years of age or older, it was more than 20%. The occurrence of clear hypothyroidism was found to be 0.4%. Taking into account both confirmed and undiagnosed cases, the presence of obvious hypothyroidism was found to be 0.37% and 3.8%, correspondingly, in a European meta-analysis. An estimated 226 incidences of hypothyroidism per 100,000 people were reported per year [1].

Through the production of hormones, the thyroid gland is essential for regulating important biological processes. These hormones affect energy levels, metabolism, and the healthy operation of organs. Thyroid hormones play a pivotal role in pregnancy, influencing fetal organogenesis, growth, and the development of the brain's structure, thereby impacting cognitive abilities [2].

By considering a healthy lifestyle, the thyroid gland function plays an important role. The improper functioning of this gland leads to different health related issues like saviour hormonal disorder and metabolic syndrome [3].

Pregnancy is a period during which thyroid dysfunction can adversely affect both maternal health and fetal development [4]. In iodine-deficient regions such as Uzbekistan, pregnant women are at a higher risk of developing subclinical hypothyroidism and hypothyroxinemia [5]. Therefore, early identification of thyroid disorders during pregnancy is important. Long-term iodine deficiency, which remains a public health concern in Uzbekistan, contributes to a range of conditions known as iodine deficiency disorders (IDD). These disorders are associated not only with thyroid abnormalities but also with impaired reproductive health, adverse pregnancy outcomes, delayed child development, and reduced cognitive function [6].

Thyroid hormones play an essential role in brain development. They are involved in neurogenesis, neuronal differentiation, cell migration, synapse formation, and myelination. Insufficient hormone levels during fetal life and early childhood can impair brain maturation and may result in intellectual and neurological deficits [7].

Machine learning methods have recently been applied to thyroid disease prediction due to their ability to identify patterns in large clinical datasets. These methods can support earlier and more accurate diagnosis, assisting healthcare professionals in clinical decision-making.

Stacking is an ensemble learning technique that combines predictions from multiple base models through a meta-classifier. By integrating the strengths of different algorithms, stacking can improve classification performance compared with the use of a single model.

The objective of this study is to develop a healthcare monitoring system for thyroid disease prediction using a stacking-based machine learning framework. The proposed approach aims to improve diagnostic accuracy by combining multiple classifiers within a unified predictive model.

2. Literature Review

Chaubey et al. [8] used the Iterative Dichotomiser algorithm (ID3) and other ML algorithms. Amongst all the applied models, KNN achieved the highest accuracy of 96.875%. The study had problems of an imbalanced dataset, underscoring the need for robust strategies in handling imbalanced class distributions.

Salman et al. [9] study included approximately 1250 Iraqi individuals. Their work applied MLP which achieved an accuracy of 97.6%. The study stated the issue of imbalanced data, highlighting the challenges in achieving a balance between predictive performance and dataset equilibrium.

Rehman et al. [10] adopted the min-max method for feature scaling and explored feature selection methods such as L1 and L2. The evaluation phase incorporated metrics like the MCC and error rate, showcasing a meticulous approach in identifying the algorithmic models that exhibit better performance in the given context.

Aversano et al. [11] explored the possibility of predicting LT4 treatment trend for hypothyroid patients. Data pre-processing techniques like interpolation, discretization, balancing and normalization were done before the application of Synthetic Minority Oversampling Technique (SMOTE). Several Boosting methodologies including CatBoost and XGBoost with K-fold cross-validation were applied.

Chaganti et al. [12] used different feature selection methods for feature reduction. Furthermore, different deep learning models were employed, to accurately predict the presence of thyroid syndrome.

Rao et al. [13] directed their attention to a kaggle dataset, proposing a system that involved decision trees and the Naive Bayes theorem. The choice of features, including age, gender, T3, T4, and TSH, reflected a tailored approach to the specificities of the dataset under consideration

Duggal et al. [14] explored different techniques with ML models like SVM, Random Forest and Naïve Bayes for categorizing diseases into various classes with focus on achieving precision and accuracy

Hu et al. [15] tackled the challenges posed by a large dataset by applying gradient boosting techniques, including CatBoost, alongside a 3-layered ANN. Their approach to algorithm selection underscored a commitment to harnessing the strengths of various models in tandem.x

Arjaria et al. [16] explored the application of explainable AI in healthcare by employing SHapley Additive exPlanations (SHAP) to interpret machine learning predictions. Their work highlighted the importance of making machine learning models easier to understand and interpret, especially in healthcare .

Bhattacharjee et al. [17] applied dimensionality reduction techniques such as PCA and SVD to handle high-dimensional data. They then used a deep neural network with data augmentation and optimized layer settings to improve prediction performance.

Savcı et al. [18] tackled the challenges posed by the "curse of dimensionality" in high- dimensional data and emphasized the importance of pre-processing techniques. The study utilizes methods such as mutual information for feature selection, to filter out features before model fitting. The pre-processing involves scaling data to normalize variables, ensuring accurate prediction. Techniques like min-max normalization are employed to balance data values. Correlation feature selection is applied to identify and manage correlated features, focusing on selecting variables highly correlated with the target.

Alyas et al. [19] explored the global impact of thyroid diseases, covering hypothyroidism, hyperthyroidism, and thyroid cancer. The random forest algorithm achieved 94.8% accuracy and a specificity of 91%.

After studying the papers of different researchers in the literature, the below research gaps were pointed:

1. Much of the referred research did not deal with imbalanced dataset. A gap of methodologies for dealing with this issue and its effect on model performance was observed.
2. Feature selection methodology was done by many of the studies, but a standardized approach is still missing that can improve model performance for other medical domains.
3. Several of the studies achieved high accuracy on their data but their performance on unseen data is still unknown for real life scenarios.
4. The reviewed studies did not discuss the use of hyper parameter tuning which could result in better model performance.

3. Methodology of study

The methodology of study , illustrated by Figure-1, outlines a comprehensive workflow divided into two key components: Data procurement & Pre-processing, and Model Training & Disease Identification. During data pre-processing, tasks including handling missing values, balancing the dataset, selecting pertinent features, and scaling them were executed. Following this pre- processing stage, the training of ML models like KNN, SVM, and Random Forest is conducted utilizing the prepared training samples.The trained models undergo validation using separate test data to ensure their reliability and effectiveness in identifying diseases.

3.1 Data Collection

The study sourced its data from the UCI ML Repository. Initially consisting of 30 features, the data consisted of 3772 entries. However, some entries contained missing values, necessitating further consideration and handling of these gaps in the data. The list of all the features, along with description is given in **Table 1**.

Table 1: Name and Data type of the Features in the UCI Dataset.

Data Type	*Features*
Numerical	*Age, TSH, T3, TT4, T4U, FTI*
Categorical	*Sex, On Thyroxine, Query on Thyroxine, On Antithyroid Medication, Sick, Pregnant, Thyroid Surgery, I131 Treatment, Query Hypothyroid, Query Hyperthyroid, Lithium, Goitre, Tumor, Hypopituitary, Psych, TSH Measured, T3 Measured, TT4 Measured, T4U*

	Measured, FTI Measured, TBG Measured, TBG, Referral Source, Binary Class

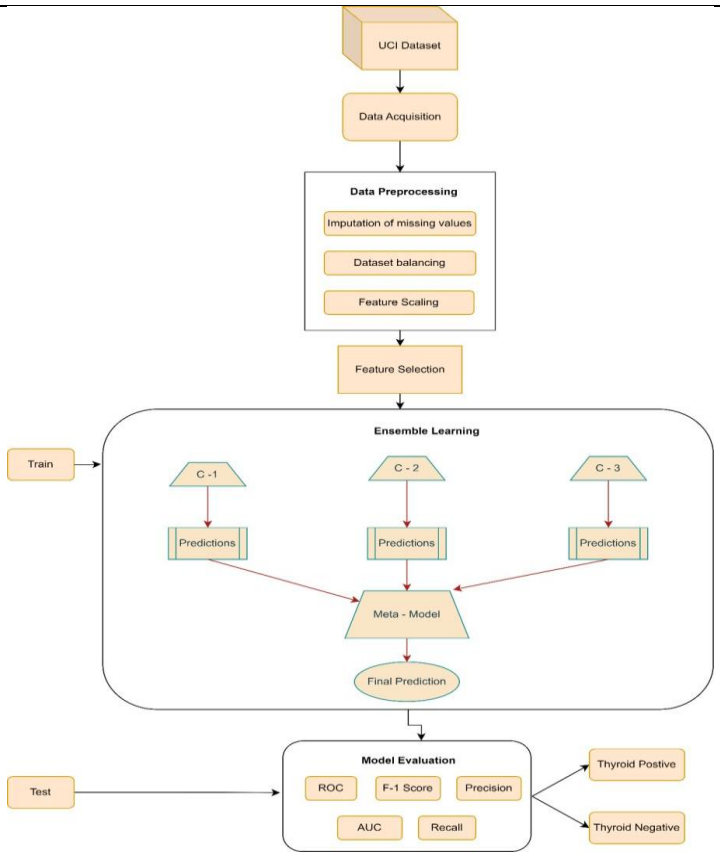

Figure 1: Proposed System Architecture

3.2 Data pre-processing

Data refining is the most critical part of model development. If the data is not processed properly, it could give a very disappointing performance. The number of missing values in all features were visualized in bar-chart as shown in **Figure 2**. In this system, we have first replaced all the missing "?" values with NaN for simplifying the dataset. Then the column "TBG" was dropped completely from analysis as it had no entries in it.

Next, the Boolean values in columns like "goitre", "tumor", etc. were replaced with 0 and 1. Then the missing values were replaced with the median, median was chosen over mean as it is not affected by potential deviations in the data, making it a more robust choice for replacing the missing values. During the analysis, it was noticed that the final output class of the dataset was imbalanced, having 3341 values of one class and only 280 values of the other class. So, to overcome this problem, an oversampling technique called SMOTE was used to increase the minority sample size.

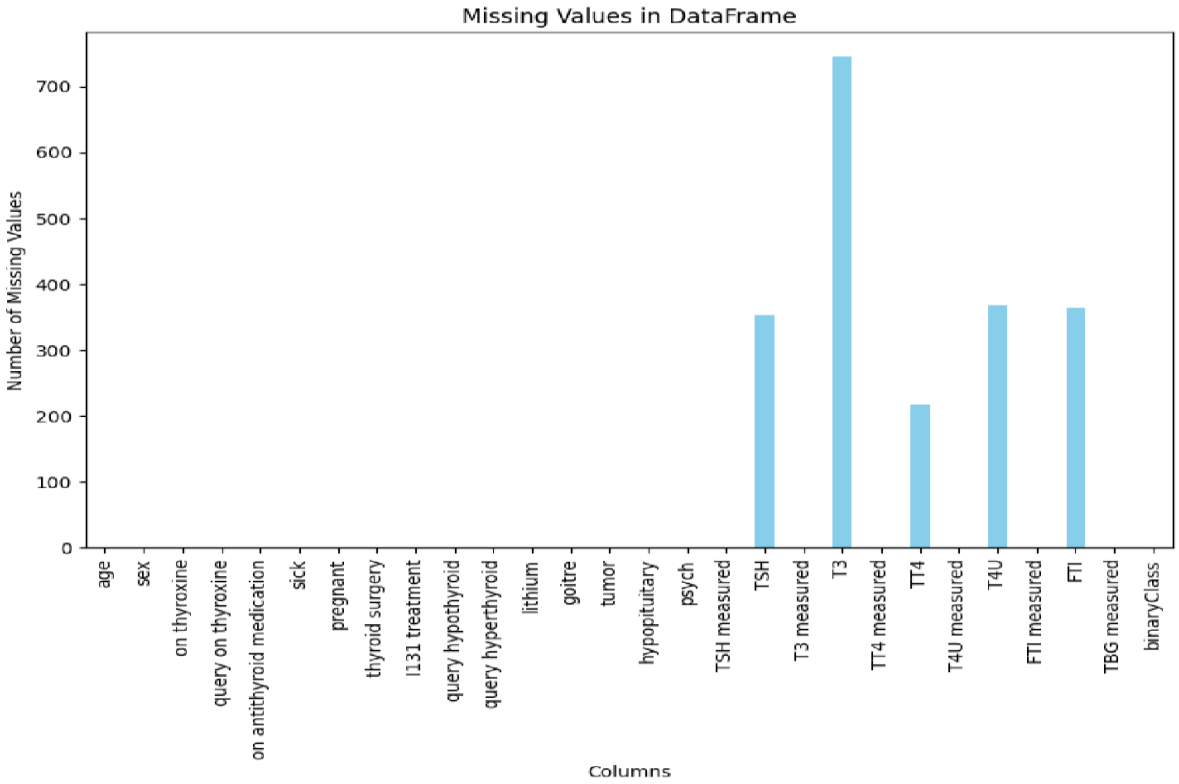

Figure 2: Number of missing values

i. Data balancing using SMOTE

When the distribution of categorical classification columns is significantly skewed, the dataset is characterized as imbalanced as shown in **Figure 3**. SMOTE, short for Synthetic Minority Oversampling Technique tackles imbalanced datasets by generating augmented instances for the minority class. This is achieved by creating artificial data points situated between existing minority class samples and their K-nearest neighbors. These new samples are randomly positioned along the lines connecting the source data points. The selection of the number of neighbors selected is contingent upon the desired level of oversampling, providing a flexible approach to rebalancing the dataset [20]. After performing SMOTE, the class distribution is as shown in **Figure 4**

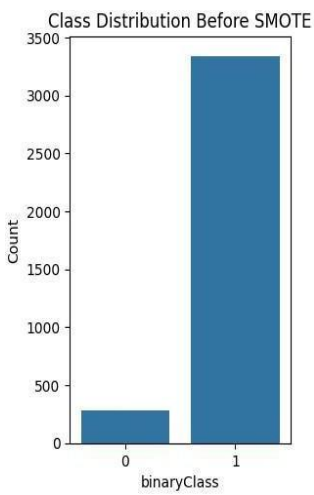

Figure 3: Imbalanced Distribution

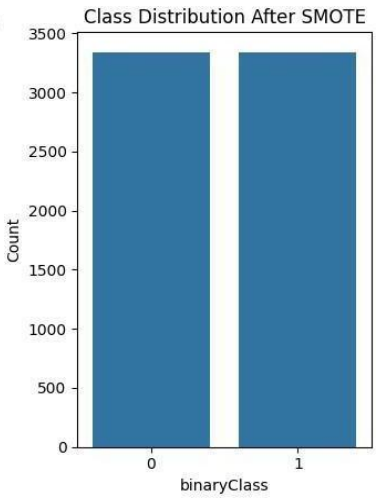

Figure 4: Balanced Distribution

ii. Feature Selection

Feature selection involves choosing of relevant features from a dataset for model application. This process eliminates less-contributing features to the end result. This methodology is done to get better performance.

a. Chi-Square Test:

Pearson's chi-square test is a non-parametric test that evaluates whether the observed frequency distribution of categorical data differs significantly from the distribution expected under a null hypothesis of independence. The test serves two primary purposes. First, it tests the hypothesis. Second, it evaluates how well the observed data fits an expected distribution, thereby determining the goodness-of-fit [21]. The Chi-square test results have been visualized as shown in Figure 5

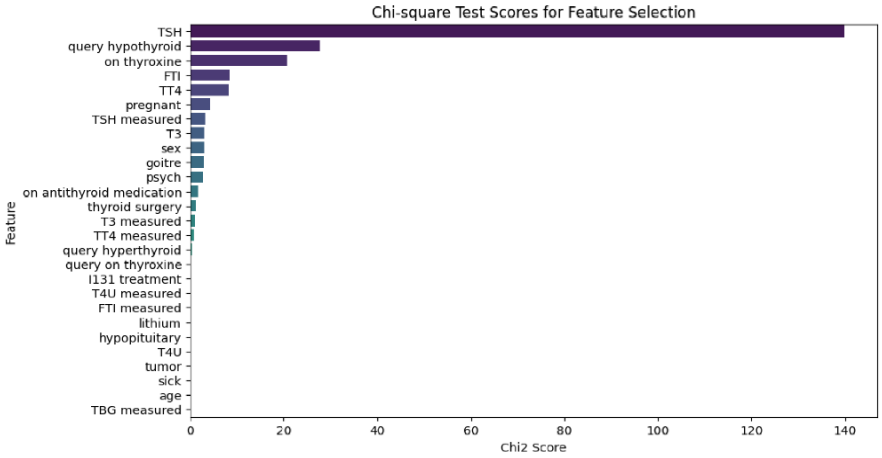

Figure 5: Chi square test scores of all features

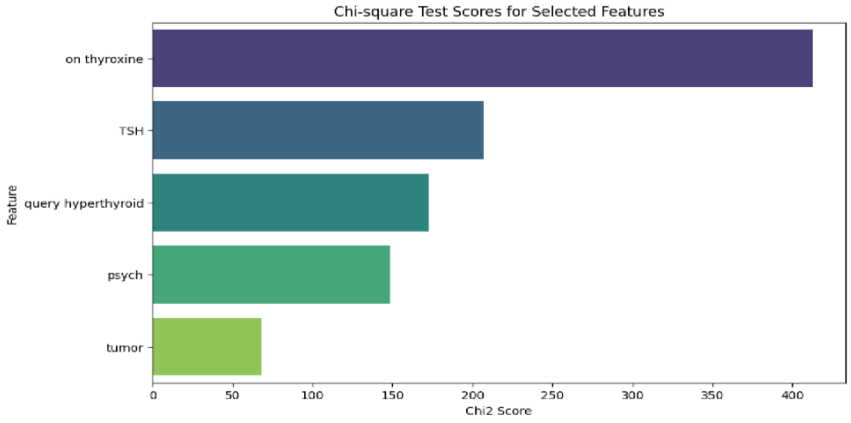

Figure 6: Chi square test scores of selected features

The expected frequencies are computed under the assumption that the null hypothesis holds true. The equation is as follows:

$$\chi^2 = \Sigma[(Observed_i - Expected_i)^2 / Expected_i] \quad (1)$$

Where:

$Observed_i$: observed frequency of the *i*th category.

$Expected_i$: expected frequency of the *i*th category.

n: total number of categories.

This statistical test offers significant understanding into the association between categorical variables and the goodness-of-fit of observed data to expected distributions.

Figure 5 presents the chi-square scores of each a feature against the dependent variable in the thyroid data set. Chi-square scores denote the level of independence of each attribute from the target variable. The highest chi-square score is of TSH, followed by FT4 and TT4. These values show the significance of the three features in predicting the dependent variable, compared to other features like T3, Age, Sex, and BMI.

Figure 6 below shows the chi-square scores associated with the selected features as predictors of the dependent variable. The selected features include on thyroxine, TSH, query hyperthyroid, psych, and tumor. On thyroxine has the highest chi-square score, which implies a strong statistical relationship with the dependent variable. It means that people taking thyroxine medicine can significantly impact the dependent variable and thus make them more significant predictors. Query hyperthyroid and TSH also exhibit medium-level chi-square scores, which indicate some association, but it is weaker compared to that of on thyroxine. Psych and tumor have low-level chi-square scores.

3.3 Stacking Based Classifier

The Stacking Classifier is a powerful ensemble learning approach that amalgamates predictions from various base classifiers to enhance overall efficiency. In this research, the Stacking Classifier is employed to further boost the predictive accuracy of our models in comparison to individual classifiers such as SVM, Random Forest, and KNN trained with Grid Search CV.

Random Forest

Random forests utilize ensemble learning, incorporating multiple decision trees $t_0, t_1, t_2, \ldots\ldots, t_n$. Each tree t_i within the forest is constructed using a random vector that is independently sampled from the same distribution for all trees, ranging from 0 to n-1. As the number of trees in the forest grows, the generalization error stabilizes. The overall prediction accuracy depends on both the individual effectiveness of each tree and their correlation. In simple terms, random forests combine several decision trees, each using randomly sampled data. This approach leads to more accurate predictions by avoiding overfitting issue [22] [30].

The importance of a feature in a random forest model is calculated by measuring how much the feature contributes to the reduction of the impurity of the trees in the forest as shown in equation 2.

$$RFf_i = \sum_{z=1}^{m\ trees} \frac{(\ Gini(parent_z) - Gini(child_z)\ )}{m\ trees} \tag{2}$$

Where:

- RFf_i is the importance of feature i in the model
- $Gini(parent_z)$ denotes the Gini impurity of parent nodes of the split
- $Gini(child_z)$ represents Gini impurity of the child nodes of this split
- $m\ trees$ stands for the number of trees in the model

Gini index calculates the probability W_i of a particular data point Di when it is being wrongly classified to a randomly selected class *C*, as shown in equation 3.

$$\text{Gini(node)} = 1 - \sum_{i=1}^{c} W_i^2 \tag{3}$$

Where:

- W_i signifies the ratio of data points in the node categorized as class i
- C indicates the number of classes

Support Vector Machines

Support Vector Machines are a type of supervised learning models extensively utilized for classification and regression tasks, celebrated for their robustness and efficacy across diverse domains [23] [31]. The model constructs

a hyperplane based on a collection of training samples to maximize the separation between data points of different categories. This marks a clear division between the classes. SVMs are proficient in addressing non-linear classification challenges through the kernel trick, which involves transforming data into higher-dimensional feature spaces as shown in equation 4.[24].

The equation for the SVM hyperplane is as follows:

$$w^T x + c = 0 \quad (4)$$

Where:

- w is the weight vector
- x is a data point
- c is the bias term

By solving the optimization problem the weight vector j and the bias term b are found as shown in equation 5:

$$0.5 * \|w\|^2 + D \sum_{i=1}^{n} (0,1 - y_i(w^T x_i + c)) \quad (5)$$

Where:

- n symbolizes the number of data points
- y_i depicts the label of the data point i
- Dspecifies the hyper parameter that controls the complexity of the model

K - Nearest Neighbour

The K-Nearest Neighbor (k-NN) algorithm is an algorithm used for classifying data according to their neighbors. Unlike other algorithms models, k-NN keeps all training points in the memory which allows new data to include without re-training the model each time. Some disadvantages are, such as memory consumption and vulnerability to noise, the simplicity and intuitiveness of the algorithm make it popular. Its performance depends on the choice of the distance metric used. (see Equation 6). [26].

$$D=(\sum_{i=1}^{b} |K_i - N_i|^q)^{1/q} \quad (6)$$

Where:

- K and N are two variables
- D is the distance
- q can be any real value, commonly set as 1 or 2. If q=1, it corresponds to Manhattan distance and if q=2, it becomes Euclidean distance.

Hyper-Parameter tuning using Grid Search CV

Grid Search Cross-Validation is a hyper parameter tuning technique which depends on the exhaustive methodology to procure the best performing model parameters. It checks every possible combination of the list of parameters provided and evaluates them by using k-Fold cross-validation.The method is a cartesian product of number of parameters hence it may require higher computational strength.[27,33].

The best parameters for each base estimator were selected through Grid Search CV, optimizing their performance. For each model, the optimal parameters were identified as shown in Table 2. These fine-tuned models were then used for Stacking Classifier, using SVM, Random Forest, and KNN as base estimators, with Random Forest serving as meta learner. This Stacking Classifier surpassed its individual counterparts across key performance metrics. Specifically, it achieved an accuracy of 99.77%, precision of 100%, recall of 99.56%, and F1 score of 99.78%. The analysis of the confusion matrix revealed minimum misclassifications.

Model	Best Parameters
Support Vector Machine	C: 10, gamma: 'scale', kernel: 'rbf'
Random Forest	min_samples_split: 2, n_estimators: 200
K-Nearest Neighbors	n_neighbors: 3, p: 1, weights: 'distance'

Table 3: Optimal Parameters selected by Grid Search CV

Where,

- C is a regularization parameter
- Gamma denotes kernel coefficient for Radial Basis Function kernel
- Kernel specifies the kernel type to be used in the algorithm
- Min_samples_split specifies the minimum number of samples required to split an internal node.
- N_estimators symbolizes the number of trees in the forest.
- N_neighbors are the number of neighbors used for classification.
- P: Power parameter for the Minkowski metric.
- Weights: Weight function used in prediction.

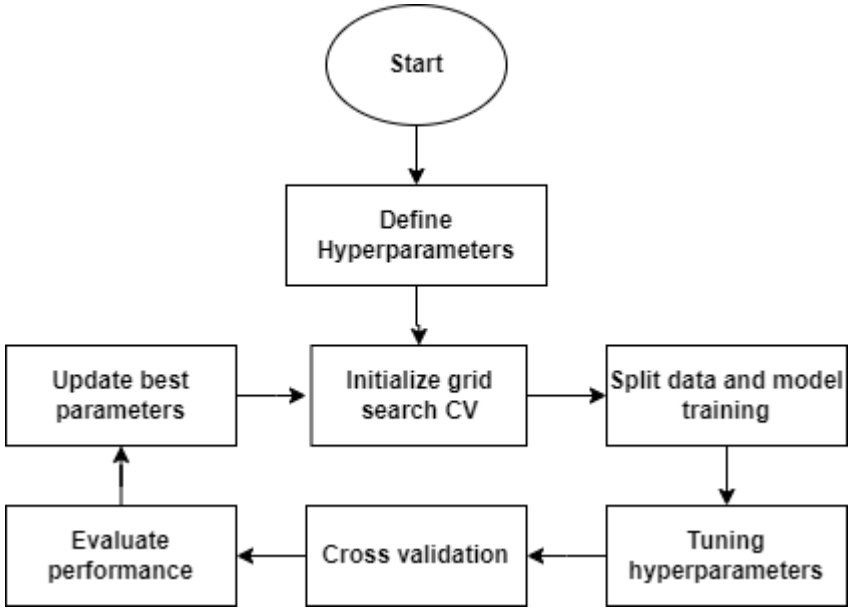

Figure 7: Hyper parameter tuning using Grid Search CV

4. Results

The best way to measure the robustness of the models is to observe and analyze the different performance parameters such as accuracy, precision, recall, F1 Score, ROC curve, Precision-Recall Curve and confusion matrix. Accuracy is simply the ratio of the number of correct predictions over the total number of predictions made. Precision is the ratio of true positives to the sum of true positives and false negatives. It basically tells how many positively predicted values are actually positive. Recall is the ratio of the true positives to the sum of true positives and false negatives. Higher Recall values means that the machine learning model is good at capturing positive instances, minimizing the negatives. F1 Score is simply the harmonic mean of precision and recall. It ranges between 0-1, higher value meaning better performance in terms of both precision and recall [28]. All these above discussed metrics formulae are shown in the following equation 7, 8, 9 and 10.

$$Accuracy = \frac{A + C}{A + C + B + D} \quad (7)$$

$$Precision = \frac{A}{A + B} \tag{8}$$

$$Recall = \frac{A}{A + D} \tag{9}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

Where:

A = True Positives, B = False Positives, C = True Negatives, D = False Negative

Receiver Operator Curve (ROC) which is a metric that gives the relation between True Positive Rate (TPR) and False Positive Rate (FPR). The Area Under the ROC gives us another measure to check the performance. AUC lies between 0 and 1 (1: perfect model prediction) and refer Figure 8,9 and 10.

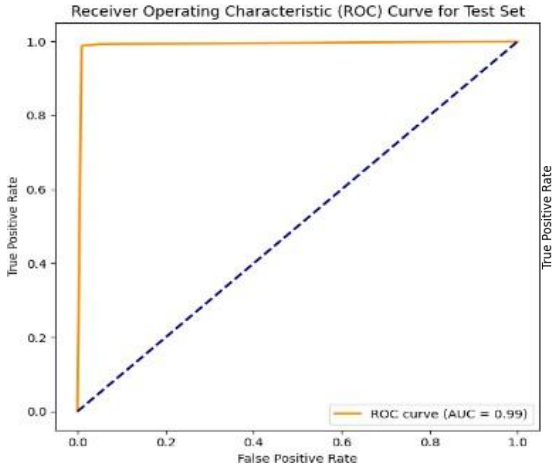

Figure 8: ROC Curve of KNN

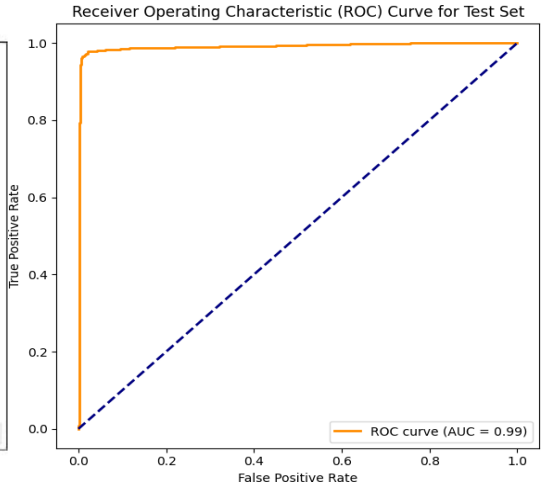

Figure 9: ROC Curve of SVM

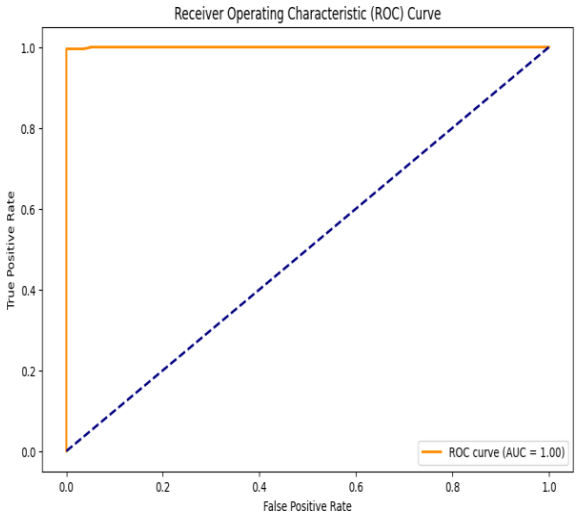

Figure 10: ROC Curve of Random Forest

Precision-Recall curve visualizes the trade-off between precision and recall. The area under the curve shows the effectiveness of the classifier. In all the 3 classifiers the score is very much close to 1. The PR-curve of all3 classifiers are shown in figure- 11, 12, 13.

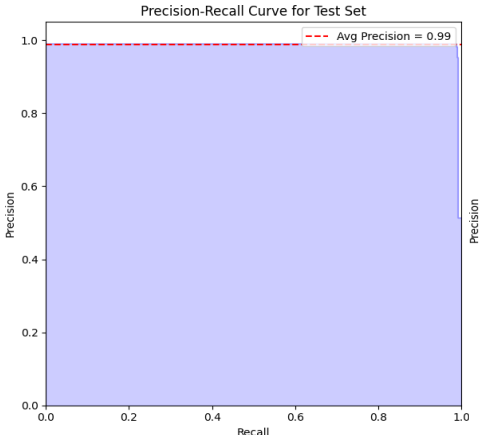

Figure 11: PR Curve of KNN

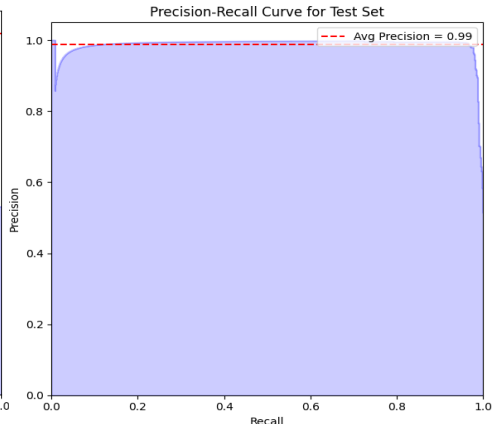

Figure 12: PR Curve of SVM

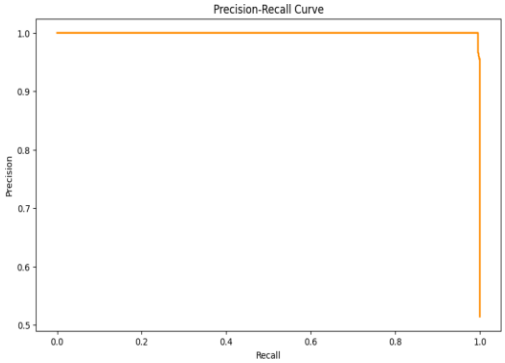

Figure 13: PR Curve of Random Forest

Table-3 presents the results for the 3 ML models that were employed before data pre-processing. KNN had the highest precision amongst the three models. However, Random Forest (RF) outperformed the rest in F1-score (RF: 0.997) and Accuracy (RF: 99.5%). SVM did competitively well yet was unable to achieve more than the rest.

Model Name	Precision	Recall	F1 Score	Accuracy
KNN	0.95	0.86	0.9	98.25%
SVM	0.74	0.86	0.8	97%
Random Forest	1	0.99	0.997	99.5%

Table 3 Results prior to data pre-processing

The application of the chi-square test and hyper-parameter tuning for thyroid prediction was done to test the performance of the models.

The best parameters for RF were (min_samples_split: 2, n_estimators: 200) and a best cross-validation score of 0.9981. Test set metrics performance, with accuracy at 0.9970, precision at 0.9985, recall at 0.9956, and F1 score at 0.9971. RF outperformed both KNN and SVM as shown in table-4.

Model Name	Accuracy	Precision	Recall	F1 Score
KNN	0.9858	0.9827	0.9898	0.9862
SVM	0.9768	0.9838	0.9709	0.9773
Random Forest	0.997	0.9985	0.995	0.9971

Stacking Classifier	0.9977	0.1	0.9956	0.9978

Table 4 Results of models with HPT parameters

The models had different performances depending on the metrics used before pre-processing. KNN showed high precision of 0.95 compared to recall, which was lower at 0.86. The model therefore performed well with high ability to accurately predict the true positives but possibly performed less in predicting true negatives. SVM had a relatively low precision of 0.74 compared to recall at 0.86. This meant that SVM performed less with high false positive prediction. Random Forest gave near perfect results for both metrics with values being 1.0 and 0.99 respectively.

In Table 5, the efficiency of all the models got better on most parameters after data preprocessing. The accuracy of KNN increased to reach 0.9902 from 0.95, which is an improvement regarding true positive classification capability. Moreover, the precision of SVM increased from 0.74 to 0.9725, which means a decrease in false positives.

All the metrics showed improvement after pre-processing methodologies were applied. Reduced false positive predictions, as indicated by the increased precision values, mean fewer instances of unnecessary medical interventions for patients who are healthy.

5. Comparative Analysis

In this section, the performance of the proposed methodology is compared with various methods discussed in the literature, as shown in Table-5 and Figure-14. The following tables prove that the reviewed papers lacked in accuracy compared to the proposed work due to the gap addressed in the Literature review.

Reference No.	**Model Name**	**Accuracy**
3	KNN	96.8%
4	MLP	97.6%
9	SVM	92.9%
6	Extra-Tree Classifier	84%
14	Random Forests	94.8%
	Proposed Work (Stacking Classifier)	**99.78%**

Table 5 Comparative Analysis

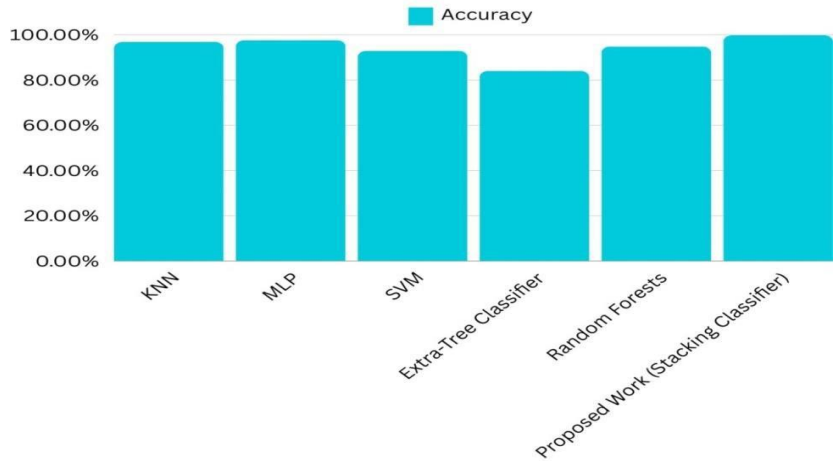

Figure-14 Comparative Analysis

6. Conclusion

In this study, a machine learning based thyroid disease identification system was developed. The study included experimentation with total features, balanced Features and reduced features. Performance analysis was done using metrices such as Accuracy, Precision, F-1 score and Recall. From the presented methodology it was observed that bagging based classifier with most correlated features from the procured data had better accuracy than the rest. The proposed system can be implemented as end-to-end pipeline for real time detection of thyroid. Medical images like CT-Scan and MR can be analyzed to be more informative.

References:

1. Chaker, L., Razvi, S., Bensenor, I. M., Azizi, F., Pearce, E. N., & Peeters, R. P. (2022). Hypothyroidism (primer). Nature Reviews: Disease Primers, 8(1).
2. Shukhratovna, N.G., Siddiqovna, T.G., Davranovna, D.A. and Nodirovna, A.S., 2022. Analysis of the thyroid status of pregnant women in the iodine-deficient region. *The American Journal of Medical Sciences and Pharmaceutical Research*, *4*(01), pp.74-78.
3. McAninch, E. A., & Bianco, A. C. (2014). Thyroid hormone signaling in energy homeostasis and energy metabolism. Annals of the New York Academy of Sciences, 1311(1), 77-87.
4. Alemu, A., Terefe, B., Abebe, M., & Biadgo, B. (2016). Thyroid hormone dysfunction during pregnancy: A review. International journal of reproductive biomedicine, 14(11), 677.
5. Rohner, F., Nizamov, F., Petry, N., Yuldasheva, F., Ismailov, S., Wegmüller, R., ... & Woodruff, B. A. (2020). Household coverage with adequately iodized salt and iodine status of nonpregnant and pregnant women in Uzbekistan. Thyroid, 30(6), 898-907.
6. Nazarpour, S., Tehrani, F. R., Simbar, M., & Azizi, F. (2015). Thyroid dysfunction and pregnancy outcomes. Iranian journal of reproductive medicine, 13(7), 387.
7. Bernal, J. (2007). Thyroid hormone receptors in brain development and function. Nature clinical practice Endocrinology & metabolism, 3(3), 249-259.
8. Chaubey, G., Bisen, D., Arjaria, S. *et al.* Thyroid Disease Prediction Using Machine Learning Approaches. *Natl. Acad. Sci. Lett.* **44**, 233–238 (2021).
9. Heuer, H. (2007). The importance of thyroid hormone transporters for brain development and function. Best Practice & Research Clinical Endocrinology & Metabolism, 21(2), 265-276.
10. Abbad Ur Rehman, H., Lin, CY., Mushtaq, Z. *et al.* Performance Analysis of Machine Learning Algorithms for Thyroid Disease. *Arab J Sci Eng* **46**, 9437–9449 (2021).
11. Aversano, L., Bernardi, M.L., Cimitile, M., Iammarino, M., Macchia, P.E., Nettore, I.C. and Verdone, C., 2021. Thyroid disease treatment prediction with machine learning approaches. *Procedia Computer Science*, *192*, pp.1031-1040.
12. Chaganti, R.; Rustam, F.; De La Torre Díez, I.; Mazón, J.L.V.; Rodríguez, C.L.; Ashraf, I. Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques. *Cancers* **2022**, *14*, 3914.
13. R. Rao and B. S. Renuka, "A Machine Learning Approach to Predict Thyroid Disease at Early Stages of Diagnosis," 2020 IEEE International Conference for Innovation in Technology (INOCON), Bangluru, India, 2020, pp. 1-4.
14. P. Duggal and S. Shukla, "Prediction Of Thyroid Disorders Using Advanced Machine Learning Techniques," 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2020, pp. 670- 675.
15. Hu, M., Asami, C., Iwakura, H., Nakajima, Y., Sema, R., Kikuchi, T., ... & Sakakibara, Y. (2022). Development and preliminary validation of a machine learning system for thyroid dysfunction diagnosis based on routine laboratory tests. Communications Medicine, 2(1), 9.
16. Arjaria, S. K., Rathore, A. S., & Chaubey, G. (2022). Developing an explainable machine learning-based thyroid disease prediction model. International Journal of Business Analytics (IJBAN), 9(3), 1-18.
17. Jha, R., Bhattacharjee, V., & Mustafi, A. (2022). Increasing the prediction accuracy for thyroid disease: a step towards better health for society. Wireless Personal Communications, 122(2), 1921-1938.
18. Savcı, E. and Nuriyeva, F., 2022. DIAGNOSIS OF THYROID DISEASE USING MACHINE LEARNING TECHNIQUES. *Journal of Modern Technology & Engineering*, *7*(2).
19. Alyas, T., Hamid, M., Alissa, K., Faiz, T., Tabassum, N., & Ahmad, A. (2022). Empirical method for thyroid disease classification using a machine learning approach. BioMed Research International, 2022(1), 9809932.
20. Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, pp.321-357.
21. Choi, S. J. (2024). Is It Possible to Precisely Monitor Thyroid Function as It Transitions from Hashimoto's to Grave's Diseases?. Science Insights, 44(6), 1393- 1395.
22. Rana, R. and Singhal, R., 2015. Chi-square test and its application in hypothesis testing. *Journal of the practice of cardiovascular sciences*, *1*(1), p.69.
23. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).
24. Cortes, C., Vapnik, V. Support-vector networks. *Mach Learn* **20**, 273–297 (1995).

25. Cunningham, P., & Delany, S. J. (2020). k-Nearest neighbour classifiers: (with Python examples). arXiv preprint arXiv:2004.04523.
26. Wang, J., Yang, L., Liu, W., Wei, C., & Shen, J. (2024). Impaired sensitivity to thyroid hormone and risk of carotid plaque development in a Chinese health check- up population: a large sample cross-sectional study. Diabetes, Metabolic Syndrome and Obesity, 1013-1024.
27. Bergstra, James & Bengio, Y.. (2012). Random Search for Hyper-Parameter Optimization. The Journal of Machine Learning Research. 13. 281-305.
28. Zhang, J., Li, J., Zhu, Y., Fu, Y., & Chen, L. (2023). Thyroidkeeper: a healthcare management system for patients with thyroid diseases. Health Information Science and Systems, 11(1), 49.
29. Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, 5(2), 1.
30. Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian journal of internal medicine, 4(2), 627.
31. Razu, M. H., Hossain, M. I., Ahmed, Z. B., Bhowmik, M., Hasan, M. K. E., Kibria, M. K., ... & Khan, M. (2022). Study of thyroid function among COVID-19-affected and non-affected people during pre and post-vaccination. BMC Endocrine Disorders, 22(1), 309.
32. Kim, K. H., Lee, J., Ahn, C. H., Yu, H. W., Choi, J. Y., Lee, H. Y., ... & Moon, J. H.(2021). Association between thyroid function and heart rate monitored by wearable devices in patients with hypothyroidism. Endocrinology and Metabolism, 36(5), 1121-1130.
33. Kong, H., & Chen, J. (2021). Medical monitoring and management system of mobile thyroid surgery based on internet of things and cloud computing. Wireless Communications and Mobile Computing, 2021(1), 7065910.