SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V11I12P129

Volume 11 Issue 12, 308-313, December 2024

Original Article

Automated Diagnosis to Predict the Thyroid Using Machine Learning Algorithms

Meenakshi Thalor[1*], Mrunal Pathak[2], Vandana Kale[3], Veena Bhende[4]

[1,2,3,4]*AISSMS Institute of Information Technology, Maharashtra, India.*

[1]*Corresponding Author : meenakshi.thalor@aissmsioit.org*

Received: 16 October 2024 Revised: 17 November 2024 Accepted: 15 December 2024 Published: 31 December 2024

Abstract **-** Thyroid disorders affect millions worldwide, necessitating early and accurate diagnosis to ensure effective management and treatment. In this context, the Thyroid Stage Prediction App represents a groundbreaking healthcare innovation. This mobile application leverages cutting-edge machine learning technologies to accurately predict thyroid stage progression, offering a proactive approach to thyroid disease management. The Thyroid Stage Estimator app features an intuitive interface that lets users access their medical history, test results, and related symptoms easily. The app then processes this information using algorithms to measure current thyroid levels and predict future growth. This predictive model is based on a large database of non-patient information and is continually updated to ensure reliability and accuracy. Key features of the app include stage estimation, personalized recommendations, reminders to consult with doctors, and encouragement of early intervention and treatment plans. The app also provides important information about thyroid health and wellness, allowing users to make informed decisions about their health.

Keywords **-** *Logistic regression, Machine Learning, Random forest, Support Vector Machine, Thyroid.*

1. Introduction

The Thyroid Stage Predictor fills this gap by using advanced machine learning [1], gaining insights from multiple patient data, and providing an intuitive and efficient platform. Early and prompt identification and effective management of these conditions are critical to ensuring the health of those affected. The Thyroid Stage Predictor Project has released a new mobile app that uses predictive technology to address thyroid health issues [2]. By providing tools that improve early diagnosis, implement effective interventions, and provide treatment recommendations to hospitals and physicians. This introduction provides an overview of the development of our app, and subsequent sections will delve deeper into its capabilities, benefits, and the technologies that support it [3].

Indeed, thyroid disorders represent one of the most prevalent health problems that impact tens of thousands globally [4]. The range of thyroid diseases is from hyperthyroidism to hypothyroidism; thus, dose-specific information becomes ineligible for the diagnosis and treatment. Thyroid disease has been the subject of a few studies, and the authors have assessed many of these studies to provide a thorough foundation for the disease classification. Tahir Alyas proposed [1] to introduce an empirical approach for classifying thyroid diseases. By examining the machine learning, random forest, KNN methods of fourteen features, including clinical characteristics and laboratory examinations, were analyzed to predict 30-day all-cause readmission. A random forest is a machine-learning algorithm that generates tens, hundreds and sometimes thousands of decision trees [2]. The study also found that detecting the stage of thyroid disease early is important, as failure in thyroid hormone production can be either excessive or inadequate.

Early detection and treatment can help prevent serious problems. This study uses Bayesian network architecture to achieve good results in classifying thyroid tumors. Bayesian network is a graphical representation of the dependency of variables. The study found that the Bayesian network framework can classify thyroid tumors into different types. The research also highlighted overfitting as a potential problem in machine learning. Overfitting occurs when a model learns well on training items and fails on new items. The PGN algorithm is a regularized processing algorithm that helps prevent overload used to reduce the overfitting of the parameters. Ritesh Jhal worked on [3] enhancing accuracy in thyroid disease prediction: A step towards improved health. This study employed a method to boost the precision of thyroid disease predictions. Hierarchical structure is a way of organizing data into a hierarchical structure. The study found that criteria can be more easily identified and removed from the data, thus helping to increase the accuracy of predicting thyroid disease. Research also shows that overlapping symptoms of thyroid disease can make classification difficult.

The study also found that the proposed system has social problems and user compatibility issues. Further research is needed to address these limitations. Pradeep Isawasan proposed [4] a regression approach for thyroid disease prediction while comparing crossing-over approaches and multivariate analysis. The Methodologies used to help improve the accuracy are linear regression, which is the prediction of unknown data; logistics regression, which estimates model parameters; and LOOCV, which is observations as a validation. The limitations of this study are the limited exploration of multivariate combinations and the comparison of the performances of different models based on different combinations. Md Riajuliislam proposed [5] research methodologies such as Data mining, Feature selection, Classification, and Recursive feature selection, limitations are not specified that distinct feature selection techniques or diverse classification techniques are used, and sample data size is not mentioned. Lerina Aversano proposed [6] classifiers to predict thyroid disease treatment. K-cross validation was used to split the data samples. Threats to validity were identified as constructive, internal, and external.

Riajuliislam studied [5] early-stage prediction of thyroid disease (Hypothyroid) using feature selection and classification techniques A. K.P and J. V. B. Benifa proposed [7] a neural network-based model for thyroid disease prediction, which is helped to increase the accuracy. R. Chaganti proposed [8] a system to predict thyroid disease using selected features and machine learning techniques.G. Chaubey proposed using machine learning to predict thyroid disease more effectively [9]. R. Banu proposed using a Random Forest distribution model and SVM to predict thyroid diseases [10]. This study shows that an integrated random forest and support vector machine can improve the accuracy and robustness of the model and provide a reliable method for classifying thyroid diseases. L. Aversano proposed using machine learning to predict thyroid therapy [6]. By identifying a patient's unique characteristics, these methods can help improve outcomes and adapt thyroid disease management strategies.

2. Research Methodology

Figure 1 explains the system workflow visually. First, the collected data will be stored in the dataset. Then, after prepossessing, the exploratory data analysis is performed to understand the dataset. Feature extraction utilizing Random Forest and Sequential Feature Selectors identifies crucial predictors. Finally, SMOTE oversampling tackles class imbalance. This systematic approach ensures data quality, feature relevance, and model robustness, laying a solid foundation for accurate thyroid disease prediction.

The first spitting operation is done to train the model. Then, testing data is used to evaluate the model build using a training dataset. The dataset of thyroid patients is required to construct a machine learning model. The Kaggle Machine Learning website provided the Thyroid dataset.

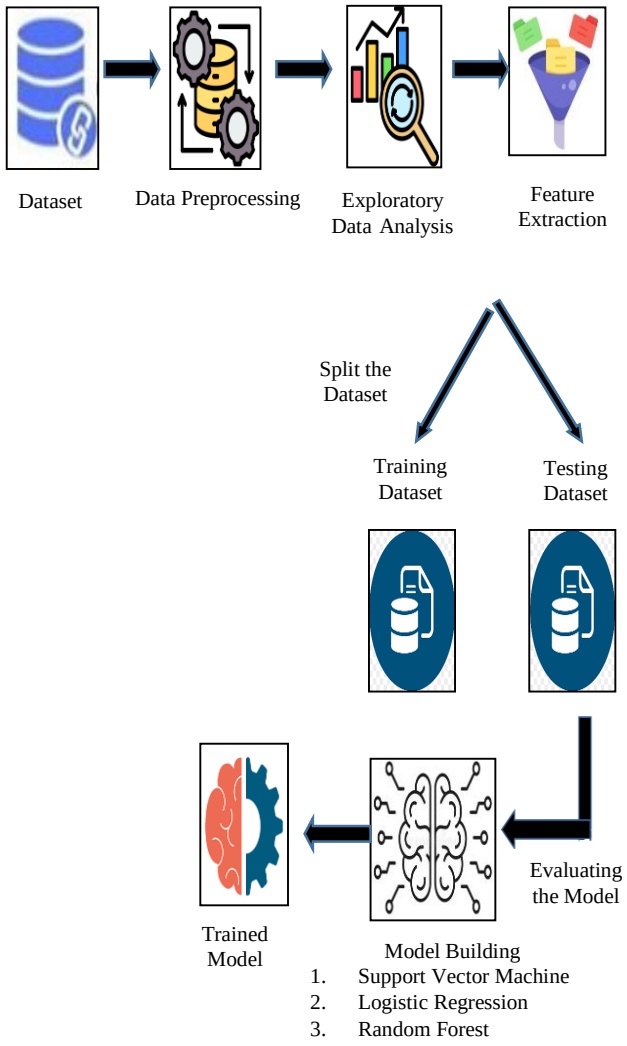

Fig. 1 System architecture of thyroid diagnosis application

2.1. Data Collection

During this work, a data set of 2799 patients (Male:860 and Female:1939) is collected from Kaggle. The Thyroid Disease Report on Kaggle is a compilation of information about thyroid disease, including features such as age, gender, thyroid hormone levels (e.g. TSH, TT4), drug use, and other treatments. This data is often used in machine learning and data analysis projects to discover patterns, patterns, and predictive models related to thyroid disease that can help with early detection, diagnosis, and treatment planning. This data includes categorical variables (e.g. medication status, gender) and numerical variables (e.g. age, hormonal strata), making it amenable to many advanced techniques such as coding categorical data and scaling numerical features.

2.2. Data Cleaning and Preprocessing

In our thyroid diagnosis project, data cleaning is a critical phase in the preprocessing stage. This involves identifying and

correcting errors in the dataset to bolster its reliability and quality. Implementing robust data cleaning techniques is essential for refining the dataset, ensuring more accurate predictions in thyroid diagnosis. This meticulous process not only eliminates noise but also establishes a solid foundation for subsequent stages, contributing to the overall robustness of our machine learning model. The following steps are carried out during the data cleaning and preprocessing stage.

1. Handling missing values
2. Dividing the dataset into the training and test set
3. Encoding categorical data
4. Handling imbalanced data
5. Exploratory data analysis

For handling imbalanced data in the data-set. SMOTE algorithm is used to overcome oversampling [12]. Exploratory Data Analysis (EDA) provides a basic understanding of the thyroid patient's data and distribution. Different plots like Count-plot, Histogram plot, Pie plot, and Pandas Profiling are used to get insight into the data and uncover the relationship between different parameters of the dataset.

2.3. Feature Extraction

Feature extraction is essential for analyzing high data or improving model performance [8]. Sequential Feature Selection (SFS) is done forward, meaning the features are added individually to evaluate their impact on the model. To ensure the quality of the analysis, 5-fold cross-validation was used in the selection process. Finally, the SFS is suitable for repeated training data and different targets based on the accuracy score as a selection process to determine the most important data for the predicted target variables using random forest distributions.

The Spearman's rank correlation is applied to find the correlation between the features. Spearman's rank correlation coefficient (ρ) measures the correlation between two variables. Convert the data for both variables into ranks. For example, rank can be from 1 to n for data points based on their values. If there are ties (same values), assign the average rank for those tied values. For each pair of data points, calculate the difference between the ranks of the two variables. Square each of these differences and add up all the squared differences.

After feature extraction, these selected features were used in model building to get exact results and high accuracy. Feature extraction is the most significant step in any machine learning project [12, 13].

2.4. Model Building

Splitting data into appropriate subsets is a fundamental step in preparing a dataset for machine learning [14, 15]. The primary goal is to have separate training, validation, and testing sets. A common approach is the 70/30 or 80/20 split, where most data is used for training and the rest for validation and testing. This paper divides the dataset into 80/20 ratios for training and testing.

2.4.1. Random Forest

The dataset has many features, so divide them into groups. One group had a thyroid, and the other had no thyroid after the thyroid detection test. The random forest algorithm generates several decision trees during training and aggregates their predictions to make final outcomes.

Pseudo code for Random Forest:

1. Initiating a random forest with n_trees=100, max_features = 5
2. Create an empty set.
   Select the best remaining features one by one with j = feature value
   If j < j+1
   Update new features in the set
   repeat the process to get a set of selected features
3. for (i=0;i<=10;i++)
   Evaluate the model
   Classify the output in the class (thyroid detected or not)
4. Train the random forest using only the selected feature.

In the above table, the model added one feature in each iteration to build decision trees and conclude the model's accuracy. This went on to 10 features in total in the last iteration. Random forest handles missing data points effectively. It ranks features based on importance while dealing with hyperparameters, which is complex and time-consuming.

2.4.2. K-Nearest Neighbor (KNN)

This model calculates the distance between a new patient's data point and existing patients in the dataset. It then identifies the 'k' closest data points, or neighbors, and determines their most common classification. The majority class among these neighbors is assigned to the new patient, who could be thyroid present or thyroid absent.

Pseudo code for K-Nearest Neighbor

1. Classify(M, N, n) M=training data, N=Class names of M, n=unknown samples
2. for i=1 to I do
   Compute Euclidean distance(Mi, n)
   End for
3. Compute set I containing indices for the k smallest distance d(Mi,n)
4. Return
   Majority labels for {Ni where i belongs to I}
   The new patient is assigned to the thyroid present or absent class based on the majority value.

KNN is easy to implement and flexible for multi-class classification but sensitive to outliers and noisy data in the thyroid patients data-set.

2.4.3. Naive Bayes Classification

Following the method for thyroid detection using prob abilistic models and assumption of independence.

Pseudo code for Naive Bayes

1. Extract the training dataset
2. Calculate the prior probabilities corresponding to each class of thyroid patients' dataset
3. Calculate the Likelihood probabilities
   For all the classes
   For all the features
   a. Extract all the features
   b. Extract the data points where the Y-value is the given class
   c. Fit the normal distribution to features
   d. Calculate the mean and standard deviation for that particular feature
   e. Place the values of mean and standard deviation in the probability density function
   End
   End
4. Calculate Posterior probabilities for all the classes of thyroid patients' dataset.
5. Calculate posterior probabilities as the product of likelihood probability probability.
6. Return the task with maximum posterior probability of concluding the respective class

Naïve Bayes works well with categorical data. Deals with irrelevant features present in the thyroid patient's data-set. It requires large data points to provide accuracy. It faced accuracy problems while handling features with fewer data points.

2.4.4. Logistic Regression

This model is used to predict the probability of a binary outcome (e.g., whether a patient has a thyroid condition or not), representing the probability of the positive class (patients having thyroid) as 1 and the negative class (patients having no thyroid) as 0.

Pseudocode for Logistic Regression

1. Instantiate the Logistic Regression object using the parameters (random_state=0, max_iteration=10). Ensures that the results are reproducible by initializing the random number generator with a fixed seed. Limits the number of iterations the algorithm will perform while optimizing the model. This parameter prevents excessive computation, though in practice, one can increase this for better convergence.
2. Weights are allocated to features, and their linear combination is calculated.
3. Now, the sigmoid value is calculated. The sigmoid function transforms any real-valued input into a value between 0 and 1, making it ideal for estimating probabilities.
4. Now, by using the decision rule for classification, apply a threshold to the probability p:
   a. If $p \geq 0.5$, classify the sample as the positive class (e.g., has thyroid condition).
   b. If $p < 0.5$, classify the sample as the negative class (e.g., does not have thyroid condition).
5. The cost function is calculated, and iterations are performed to minimize its value using a gradient descent, which is used to calculate the optimum value of coefficients.
6. The model gets created with optimum results at last.

2.5. Model Testing and Evaluation

Finally, evaluations were done on all the models to determine their accuracy. The following measures are considered to determine the accuracy,precision, recall, and F1 score of Naive Bayes, KNN, and Random Forest algorithms.

True Positive (X) : Patients having thyroid getting detected positive
True Negative (Y) : Patients having no thyroid getting detected negative
False Positive (M) : Patients having no thyroid getting detected positive
False Negative (N): Patients having thyroid getting detected negative

$$\text{Precision}(P) = X / (X + M)$$

$$\text{Recall}\ (Q) = X / (X + N)$$

$$\text{F1 Score} = 2 * P * Q / (P + Q)$$

These metrics are used to evaluate the performance of classification models, with precision focusing on the accuracy of positive predictions, recall measuring the ability to identify all relevant instances, and F1 score providing a balance between precision and recall. Table 1 shows the evaluation measures of different models.

Table 1. Performance comparison of different models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Random Forest	97.76	97.0	97.0	97.0
KNN	67.68	77.0	67.0	61.0
Naive Bayes	80.56	82.0	80.0	80.0
Logistic Regression	93.7	95.0	93.0	94.0

The table compares the accuracy of different models for thyroid detection, with Random Forest having the highest accuracy at 97.76% and KNN the lowest at 67.68%. Table 2 shows the comparison between existing methodologies and proposed methodologies.

Table 2. Comparison between existing methodologies and proposed methodologies

Sr. No.	Related Work Reference	Methodology	Accuracy (%)
1	Tahir Alyas, Muhammad Hamid [1]	Random Forest	94.8%
2	Giuseppe Mollica, Daniela Francesconi [2]	SVM, PGM, DT	96%, 78%, 96%
3	Ritesh Jhal, et al. [3]	SVM, Decision Tree	97.35%, 98.7%
4	Pradeep Isawasan [4]	Logistic Regression, Linear Regression	82.97%, 72%
5	Md Riajuliislam, et. al [5]	Navie Bayes, Random Forest, Decision Tree	89.74%, 88.46%, 87.17%
6	Lerina Aversano [6]	KNN	84%
7	Proposed Method	Random Forest, KNN, Naive Bayes, Logistic Regression	97.76%, 67.68%, 80.56%, 93.7%

The performance of the proposed system is also outstanding, considering random forest and logistic regression. Random forest is often more accurate and efficient than K Nearest Neighbor (KNN) and naive Bayes due to its ability to resolve complex patterns and interactions affected by the data. Random forest creates a series of decision trees and averages their predictions, which helps control competition and capture the relationship between features, making it robust or unaffected by noisy objects. In contrast, KNN can be slow and sensitive to parameter values and noise, while Naive Bayes relies on the assumption of independence, which may not be valid for thyroid data with interacting features.

3. Conclusion

Thyroid issues are crucial due to their prevalence and potential impact on health. This paper proposed a multifaceted approach to predict the thyroid using KNN, Naive Bayes, Random Forest and Logistic Regression. Therefore, combining random forest and automatic selection often leads to better performance in thyroid research. Using the above three algorithms, one can predict whether a patient has thyroid issues based on their health parameter values. This could aid early detection and prompt medical intervention. This personalized prediction can assist in tailoring treatment plans and monitoring. KNN, Naive Bayes, Logistic Regression, and random forest have 67.68%, 80.56%, 93.7%, and 97.76% accuracy, respectively. Therefore, it is concluded that Random Forest is more efficient. In the future, an ensemble-based approach can provide a more robust application.

References

[1] Tahir Alyas et al., "Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach," *BioMed Research International*, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[2] Giuseppe Mollica et al., "Classification of Thyroid Diseases Using Machine Learning and Bayesian Graph Algorithms," *IFAC PapersOnLine*, vol. 55, no. 40, pp. 67-72, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] Ritesh Jhal, Vandana Bhattacharjee, and Abhijit Mustafi, "Increasing the Prediction Accuracy for Thyroid Disease: A Step towards Better Health for Society," vol. 122, no. 2, pp. 1921-1938, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[4] Song-Quan Ong, Pradeep Isawasan, and Khairulliza Ahmad Salleh, "Regression Study for Thyroid Disease Prediction: Comparison of Crossing-Over Approaches and Multivariate Analysis," *2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, IPOH, Malaysia, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[5] Md Riajuliislam, Khandakar Zahidur Rahim, and Antara Mahmud, "Prediction of Thyroid Disease (Hypothyroid) In Early Stage Using Feature Selection And Classification Techniques," *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, Dhaka, Bangladesh, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] Lerina Aversano et al., "Thyroid Disease Treatment Prediction with Machine Learning Approaches," *Procedia Computer Science*, vol. 192, pp. 1031-1040, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[7] Anu K.P, and J. V. Bibal Benifa, "A Comprehensive Analysis Using Neural Network-Based Model for Thyroid Disease Prediction," *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, Trichy, India, pp. 72-78, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] Rajasekhar Chaganti et al., "Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques," *Cancers*, vol. 14, no. 16, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Gyanendra Chaubey et al., "Thyroid Disease Prediction Using Machine Learning Approaches," *National Academy Science Letters*, vol. 44, no. 3, pp. 233-238, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[10] R. Banu, "Classification Model Using Random Forest and SVM to Predict Thyroid Disease," *International Journal of Scientific & Technology Research*, vol. 9, no. 2, pp. 1680–1685, 2018. [Google Scholar]

[11] Amulya.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. [CrossRef] [Google Scholar] [Publisher Link]

[12] A K Aswathi, and Anil Antony, "An Intelligent System for Thyroid Disease Classification and Diagnosis," *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Coimbatore, India, pp. 1261-1264, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[13] K. Shankar et al., "Optimal Feature-based Multi-Kernel SVM Approach for Thyroid Disease Classification," *The Journal of Supercomputing*, vol. 76, pp. 1128-1143, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[14] Priyanka Duggal, and Shipra Shukla, "Prediction of Thyroid Disorders Using Advanced Machine Learning Techniques," *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, pp. 670-675, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[15] Poonam S Jadhav, and Punashri M Patil, "Machine Learning Algorithms and its Applications: A Survey," *International Journal of Technology Engineering Arts Mathematics Science*, vol. 1, no. 1, pp. 28-31, 2021. [Google Scholar] [Publisher Link]