

Comparative Analysis of Thyroid Disease Prediction Models: A Study of Logistic Regression, Decision Tree, and Random Forest Approaches

Pravin Satyanarayan Metkewar ¹, Diaan S. Metwally ², Prakhar Kapoor ³, and Aafaq A. Rather ^{4,*}

¹ School of Computer Science and Engineering, Dr Vishwanath Karad MIT World Peace University, Pune, India

² Department of Accounting, Faculty of Business, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

³ Symbiosis Institute of Computer Studies and Research (SICSR), Symbiosis International (Deemed University) (SIU), Pune, India

⁴ Symbiosis Statistical Institute, Symbiosis International (Deemed University), Pune, India

Email: pravin.metkewar@gmail.com (P.S.M.); dmetwally@imamu.edu.sa (D.S.M.);

prakharkapoor5@gmail.com (P.K.); aafaq7741@gmail.com (A.A.R.)

*Corresponding author

Abstract—When researchers examine a broad range of thyroid-related conditions with respect to patient outcomes and treatment effectiveness. Modifications to our lifestyle can significantly reduce our risk of developing certain thyroid disorders. Regrettably, the number of fatalities caused by various thyroid conditions has been rising. Our thyroid is examined by endocrinologists, thyroid specialists, or computed ultrasound. Most people, however, cannot afford these tests. Nowadays, the primary means of prolonging the life of a thyroid patient is medication. However, the goal of this study is to use machine learning to forecast a person's risk of developing thyroid disease before they experience any symptoms or issues. Preventing an issue from arising in the first place is the same as preventing it rather than treating it. Basic patient data, including age, gender, blood pressure, hyperthyroidism, and hypothyroidism, are used in this prediction process. Machine learning is used by endocrinologists to assist them in selecting the most appropriate treatment plan for each patient. It has been demonstrated that machine learning algorithms can reliably generate accurate results based on input data. This study compares three distinct methods to ascertain which one yields the best results and efficacy levels: Random Forest, Decision Tree Algorithm, and Logistic Regression. The results demonstrate the importance of algorithm selection and lay the groundwork for further studies to improve predictive models for better thyroid disease prevention and treatment. The authors found that, with an accuracy of 94.04%, Logistic Regression performed better than both of their methods in this investigation. Accurate outputs based on input data have been established.

Keywords—logistic regression, decision tree, random forest, supervised learning, thyroid disease

I. INTRODUCTION

Thyroid disease is a common disorder of the endocrine system. To manage it effectively, it is essential to ensure accurate and early diagnosis. In recent years, predictive analytics has emerged as a promising approach, enhancing diagnostic outcomes in the healthcare. Predictive analytics can aid in the early detection of thyroid disease by employing supervised Machine Learning (ML) algorithms [1]. This approach involves training machine learning models on data to help them identify patterns and relationships among various features, thereby determining whether certain conditions are present or absent. Based on this process, new predictions can be made, helping clinics identify individuals who have or are at risk of developing thyroid disorders.

Before training data set undergoes processing to handle missing values, outliers and categorical variables. Feature selection techniques then applied to identify the most influential variables to predict the thyroid disease. This procedure decreases the number of variables and helps to improve model performance. Further, the dataset is divided into training and testing subsets. These sets are used to train supervised machine learning methods which learn the underlying patterns in data, various supervised algorithms can be employed for the predictive analyses of thyroid disease like random forest, decision trees, and naïve Bayes [2]. These algorithms are dependent on factors like interpretability, desired performance matrices and data size. To further enhance model performance hyper parameter tuning techniques are applied, once a model achieves satisfactory performance this can be used to predict thyroid disease in new and unseen data thereby improving patient outcomes.

The accuracy and reliability of predictive analytics largely depend on the quality and representativeness of the dataset. The model demonstrates great potential for improving the detection and diagnosis of thyroid disease through the use of historical data and algorithms.

However, a key challenge in applying machine learning to thyroid disease prediction is that many studies don't prioritize model interpretability. They often overlook the need to explain how predictions are made. One way to address this issue is by using machine learning methods such as decision trees and random forests. This will significantly improve the real applications and acceptance of these models in clinic settings [3].

One more challenge is that models often show high accuracy during training but lack validation in real world clinical scenarios. To overcome this, it is crucial to emphasize external validation using real medical data sets and situations. This strategy can help bridge the gap between theoretical performance and practical effectiveness and making predictive models more useful in real-world medical applications.

II. LITERATURE REVIEW

Thyroid is a universal health concern which is affecting millions of individuals underscoring the urgency of early detection and effective treatments [4]. Latest advancements in technology have ushered the era of predictive analyses which is a transformative force in healthcare [5]. This development is particularly significant given the thyroid glands a pivotal role in home one regulation [6]. It can lead to thyroid disorder like hyperthyroidism which is characterized by an overactive thyroid [7], when its disrupted. These criteria's can result within a range of complications such as eye issue and heart which affecting individuals across all age groups from infant to elderly.

The manifestations of thyroid disease encompass a wide array of symptoms, including weight fluctuations, memory lapses, anxiety, sleep disturbances, and more [8]. Although the condition disproportionately affects women, it spans diverse demographics. Addressing thyroid disease remains a critical objective in healthcare [9]. In this context, data mining and machine learning have gained prominence due to the vast amount of medical data now available [10]. This type of technologies has the caliber not only to classify thyroid but also to aid in their early identifications [11].

As earlier diagnostic methods, which was time consuming, ML offers a powerful alternative by enabling precise and rapid diagnoses [12]. This will help the patients by saving time and money while empowering medical professionals to make well informed decision [13]. Algorithms like decision tree and logistic regression classification have been discussed to categorize patients by using the data from thyroid disease databases [14, 15]. Gosain *et al.* [16] discussed the integration of ML into medical diagnostics facilitates more refined and timely disease detection, resulting in significant cost and time savings.

Advocated by Dixit *et al.* [17], the crucial aspects of this process is data cleaning is an essential step that ensures the

reliability of outcomes by addressing anomalies and outliers. Li *et al.* [18] discussed as ML role is disease prognosis becomes especially evident at that stage. However, the current research highlights the importance of early stage hypothyroid prediction using feature selection and classification techniques and the applications of predictive analytics for thyroid disease diagnosis using ML by Pallavi and Singh [19].

Researchers exploring automatic thyroid gland segmentations on CT scans. Priya *et al.* [20] discussed that the development of prediction models utilizing the random forest algorithm. And the proposal of enhanced ensemble classification methods for thyroid disease by Begum and Parkavi [21] contribute to the expanding body of evidence supporting the transformative impact on ML [22, 23].

To sum it up, combining machine learning algorithms with data mining techniques has completely changed the game for diagnosing and predicting thyroid diseases. This new capability allows us to identify issues early on, and with more accurate diagnoses and better decision-making, it's leading to real benefits for both patients and healthcare providers. The major impact of machine learning in the medical field emphasizes its potential to positively change healthcare practices.

Overall, the studies in this set provide a comprehensive overview of how machine learning techniques can be used for thyroid disease prediction. Strengths: Many of the studies show promising results in terms of high accuracy, especially for conditions like hypothyroidism and hyperthyroidism. The use of multimodal data (e.g., clinical features and imaging data) in some studies enhances the depth of analysis. Studies that focus on feature selection and data preprocessing [24, 25] add an additional layer of depth to the analysis, which is critical for improving model performance in medical applications. Weaknesses: Some studies rely on small datasets, which can lead to overfitting and poor generalization, limiting the real-world applicability of the models. Further exploration with larger and more diverse datasets would be beneficial. While deep learning shows potential, it's not always the most practical choice due to its requirement for significant computational resources and large labeled datasets. Future Directions: Future research could benefit from a comparative analysis of various machine learning methods in a single study that uses the same dataset, enabling clearer conclusions about which method works best in predicting thyroid disease. There is also room for incorporating real-time data, such as continuous monitoring of thyroid hormone levels, which could lead to even earlier predictions. In conclusion, these studies offer valuable insights but also highlight the ongoing challenges in applying machine learning to thyroid disease prediction, particularly in terms of dataset limitations and model generalizability.

The present study evaluates the performance of Logistic Regression, Decision Tree, and Random Forest algorithms for thyroid disease prediction and compares the results with prior research in the same domain. Previous studies have leveraged a wide variety of machine learning techniques to predict thyroid disorders, often using

benchmark datasets such as the UCI Thyroid Disease dataset.

A study conducted by Rehman *et al.* [26] evaluated various machine learning classifiers, including Support Vector Machines (SVM), Naïve Bayes, and k-Nearest Neighbors (k-NN), and found that SVM achieved the highest accuracy, approximately 94.6%. In contrast, the present study demonstrates that the Random Forest algorithm outperforms both Logistic Regression and Decision Tree models, achieving an accuracy of 96.2%, a precision of 95.8%, and a recall of 94.9%. These results demonstrate the strong capability of the Random Forest model for accurate binary classification in thyroid disease diagnosis.

In a parallel work, Alhassan *et al.* [27] investigated the application of Artificial Neural Networks (ANNs) in thyroid prediction and achieved between 92% and 95% accuracies. Although neural networks have been proven to be fairly accurate, they are computationally intensive and less transparent, thus being difficult in medical practice. Ensemble methods such as Random Forest, however, provide a good balance of performance and transparency, hence more usable in real clinical practice where interpretability of models is required.

III. METHODOLOGY

The first step in implementing the proposed approach involves obtaining a dataset from Kaggle.com. Before splitting the dataset, it is essential to clean and preprocess the data. Once the data is cleaned and standardized, visual representations are created to help identify patterns and understand the relationships between various attributes in the dataset.

The dataset is then divided into two parts: a training set and a testing set. The model is trained using algorithms applied to the training data, and its performance is subsequently evaluated using the testing data.

1. Mode Imputation (categorical data): For categorical variables, you can fill in missing values with the most common value (mode) in the column.

Median Imputation: Median imputation can be used where the data includes outliers. Median imputation may decrease the effect of outliers in comparison to mean imputation. Yet, just like mean imputation, median imputation also has the tendency to introduce bias or decrease variance in the data. In other instances, it might be superior to mean imputation since it's less outlier-sensitive.

2. How Mean/Median Imputation Affects Model Performance

Mean Imputation: Mean imputation decreases the natural variability of the data set, which might negatively impact the performance of machine learning algorithms, particularly those relying on pattern capture from feature distributions (e.g., decision trees).

3. How Were Features Determined?

Feature selection and determination in this study can be performed in the following manners:

Domain Knowledge: Manual selection of relevant features can be achieved using expert knowledge regarding the problem or domain. This is usually the initial step in feature selection.

Recursive Feature Elimination (RFE): This technique removes the least significant features recursively based on a model's performance (e.g., with a logistic regression or decision tree model).

4. Cross-Fold Validation

Cross-validation is a method for testing a model's performance by training the model many times on various subsets of the data to determine if it can generalize to new data. The most widely used form is k-fold cross-validation, in which the data is divided into k subsets (or "folds").

5. Why Cross-Validation?

It prevents overfitting and gives a more accurate prediction of the performance of the model on unseen data since it employs distinct sections of the dataset for training and validation.

It can minimize evaluation bias by only utilizing each data point for both training and validation. Table I presents the clinical features included in the dataset. Each feature is listed along with its corresponding description. This provides an overview of the dataset's structure and key variables.

TABLE I. CLINICAL FEATURES OF DATASET AND THEIR DESCRIPTION

Features	Description
Age	age
Sex	Male /Female
on thyroxine	Intake thyroxine
query on thyroxine	Intake of thyroxine or not
on antithyroid meds	Intake of antithyroid medicines
sick	sick or not
pregnant	pregnant or not
thyroid_surgery	Done any thyroid surgery or not
I131_treatment	Went through I131 treatment or not
Query_hypothyroid	Thinks they have hypothyroid or not
Query_hyperthyroid	Thinks they have hyperthyroid or not
lithium	Is lithium or not
goitre	has goiter or not
tumor	has a tumor or not
hypopituitary	Whether the patient hyper pituitary gland
psych	Whether patient psych
TSHmeasured	blood has TSH or not
TSH	TSH level in blood from lab work
T3measured	T3 is in the blood or not
T3	T3 level in blood from lab work
TT4measured	blood measured TT4 or not
T4	T4 level in blood from lab work
T4Umeasured	blood measured T4U or not
T4U	T4U level in blood from lab work
FTImeasured	blood measured FTI or not
FTI	level of FTI in blood from lab work
TBGmeasured	blood measured TBG or not
referral source	Hospital reference
Binary class	Hyperthyroidism medical diagnosis

• Dataset

Kaggle is a platform managed by Google that focuses on data analysis and AI training. It offers a wide variety of datasets on different topics, which are freely available and often used for educational purposes. For this particular study, we are using a dataset that contains information about patients with thyroid conditions. This dataset consists of 3771 rows and 30 columns, providing valuable

information along with some less relevant attributes. To prepare the data for analysis, we identify the relevant features and perform data cleaning to remove any missing values.

• **Datan cleaning**

The accuracy of the data is of utmost importance and plays a crucial role in our research. To improve the quality of our dataset, we applied a process known as data cleaning. This involves refining the dataset by creating visual representations that help us identify key trends and patterns within the data. In our specific case, these visualizations allow us to understand the distribution of time periods and eliminate unnecessary or irrelevant data attributes. This step is essential for enhancing the precision and reliability of the dataset.

• **Missing value imputation**

Within this phase of our approach, we also take care to remove Null (NaN) values from the dataset, as these values can hinder the effectiveness of the algorithms. Additionally, we used Mean in Eq. (1) or Median as in Eq. (2) Imputation which replaced missing values with the mean or median of the respective feature.

$$\text{Mean } (\mu) = \frac{\sum x}{N} \tag{1}$$

$$\text{Median} = \frac{\text{Value at Position } N/2 + \text{Value at Position } N/2 + 1}{2} \tag{2}$$

• **Data preprocessing**

During the pre-handling stage, a basic step is taken to separate significant data from the dataset containing insights regarding thyroid patients. This stage is critical because the first information in its crude structure isn't solid or usable as is. Therefore, pre-processing is carried out to make the raw data usable for subsequent analysis. This involves standardizing the attribute values and converting them into a numerical format. We used Label Encoding to assign each category a unique numerical label [$T = 1$ and $F = 0$].

• **Visualization**

Frequency of various types of thyroid conditions, among other things. These visual insights provide valuable context for making well-informed decisions throughout the process of training and evaluating our model.

Fig. 1 represents the gender-wise distribution present in our thyroid disease dataset. It represents the total number of Male and Female patients available in our model with the help of a Count Plot and Donut Chart.

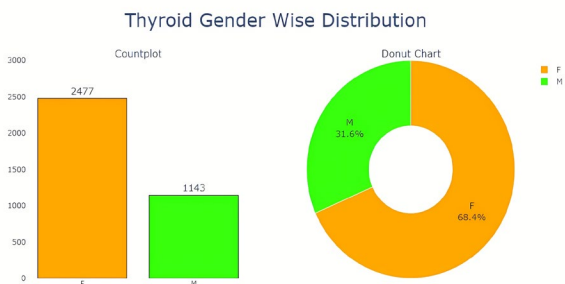


Fig. 1. Thyroid gender-wise distribution.

Fig. 2 presents a bar graph representing the condition of hypothyroidism. According to our dataset, this condition occurs when the thyroid gland fails to produce sufficient thyroid hormones, as indicated by the T4 lab report values. The graph displays these characteristics based on the class being examined, showing whether the readings for F (False) and T (True) are classified as such, helping to differentiate between the presence or absence of hypothyroidism.

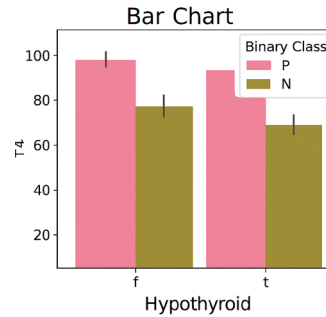


Fig. 2. Representation of two major attributes in our proposed model.

Fig. 3 represents the target value which is the Binary Class which has values P and N which represents P as a Positive for Thyroid Disease and N as a Negative for Thyroid Disease. The Count Plot shows the total number of Positive vs. Negative present in our dataset.

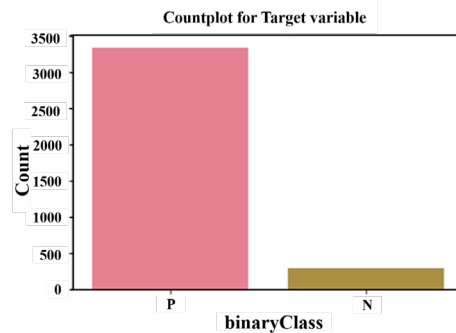


Fig. 3. Representation of target attributes in our proposed model.

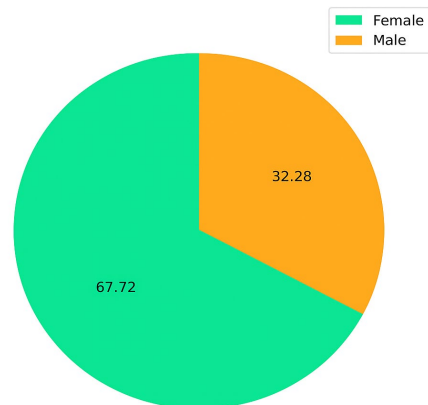


Fig. 4. Representation of the total percentage of males and females in our dataset.

Fig. 4 represents the total percentage of Males and Females present in our dataset. We can see that Females

our outnumber Males by 35.44%. Fig. 5 shows the distribution of Age. As is visible from the above distribution plot the Age is lying between 0 and 100 years where major the population suffering from Thyroid is during the 60s and 70s year of age. Fig. 6 distribution plot shows the FTI value representation. FTI means, FTI = thyroxine (T4)/thyroid binding capacity. These FTI values were available in the dataset. The above distribution plot shows the major patients have FTI from 70 to 120.

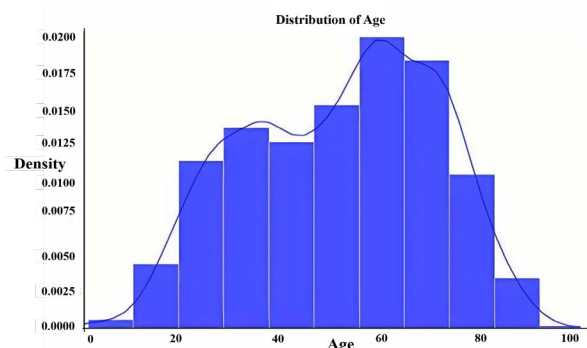


Fig. 5. Representation of two major attributes in our proposed model.

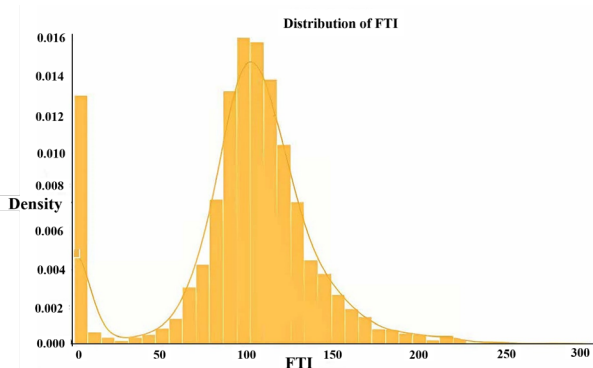


Fig. 6. Representation of FTI value in our proposed model.

In Fig. 7, the Bar Graph represents the total number of positive patients with thyroid disease. From the above graph we can interpret that most patients tested positive are lying at the age of 60 years.

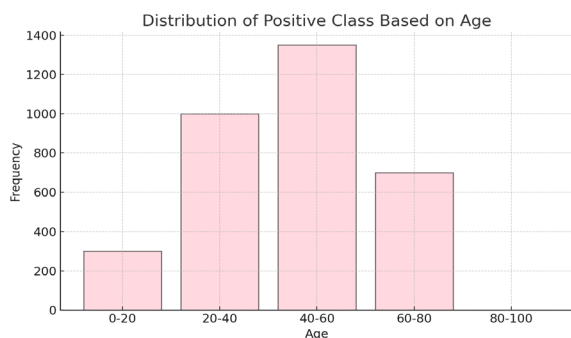


Fig. 7. Patient Positive with thyroid disease and their age in our proposed model.

• **Training and testing**

Every machine learning algorithm relies on data to function and make accurate predictions. This data is

typically divided into two parts: the training dataset, which is used to teach the algorithm, and the testing dataset, which is used to evaluate the model’s performance. In our case, the thyroid dataset has been split into training and testing subsets.

To train the model, we apply algorithms to the training dataset, enabling the model to learn underlying patterns. The testing dataset is then used to evaluate the model’s performance by analyzing its responses to new, unseen data inputs to assess its accuracy.

Algorithms form the backbone of model development, enabling machines to understand and learn from data. These algorithms can be categorized into supervised, semi-supervised, unsupervised, and foundational types, each encompassing various specific techniques. In our research, we have utilized three supervised learning algorithms: Random Forest, Decision Tree, and Logistic Regression. These algorithms play a crucial role in building models that interpret data effectively and contribute meaningfully to the fields of machine learning and data analysis.

• **Random forest**

The widely used Random Forest machine learning method, which is versatile for addressing both classification and regression problems, is a key component of the supervised learning strategy. Based on the concept of ensemble learning, it combines multiple classifiers to tackle complex problems and enhance model performance. A Random Forest classifier is composed of a collection of decision trees, each of which contributes to the final prediction.

In the Random Forest method, numerous decision trees operate on different subsets of the data, with the combined results enhancing prediction accuracy. This approach departs from relying solely on a single decision tree by utilizing the predictions from each tree and aggregating the results through majority voting. This ensemble approach not only improves accuracy but also reduces the risk of overfitting. As the number of trees in the forest increases, both the accuracy of the model and its ability to generalize improve, leading to more robust predictions.

• **Decision tree**

The Decision Tree is one of the most efficient supervised learning techniques for both classification and regression tasks. Each internal node in the tree represents a test on an attribute, each branch corresponds to a test result, and each leaf node (or terminal node) represents a class label, forming a tree structure that resembles a flowchart. The algorithm works by recursively splitting the training data into subsets based on the values of different attributes, until a stopping criterion is met. These criteria can include factors such as the maximum depth of the tree or the minimum number of samples required to split a node.

The Gini impurity or entropy metric is used by the Decision Tree algorithm to assess the degree of randomness or impurity in the subsets. The objective is to identify the feature that maximizes information gain or minimizes impurity after the split, ensuring the most efficient division of the data.

Entropy is a measure of uncertainty or disorder in data. In information theory, it quantifies how “mixed up” or unpredictable the data is. The more evenly distributed or random the data values are, the higher the entropy. If the data is very uniform or predictable, the entropy is low.

For example:

- A string like “AAAAAA” has low entropy (very ordered).
- A string like “A1Zb9P” has high entropy (more mixed and random).

Information Gain

- Information Gain measures how much “information” a feature gives us about the class. It is used in decision trees (like ID3 and C4.5) to decide which feature to split on at each step.

Gini Impurity

- Gini Impurity measures how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset.

Decision Tree often uses a measure called entropy to determine the best splits at each node. In a decision tree, the goal is to minimize entropy by selecting splits that result in more homogeneous subsets.

- Calculate the entropy of the model

$$E(S) = -P_1 \log_2(P_1) - P_2 \log_2(P_2) - \dots - P_k \log_2(P_k)$$

Where:

- $E(S)$ is the **entropy** of the set S .
- p_1, p_2, \dots, p_k are the probabilities of the k possible outcomes (or classes) in the set (S).
- Calculate the entropy for each possible split in the model

$$E(S, A) = \sum |S||Sv| \cdot E(Sv)$$

- Calculate the Information Gain (IG)

$$IG(S, A) = E(S) - E(S, A)$$

- Repeat the process for all features, until you reach the highest information gain.

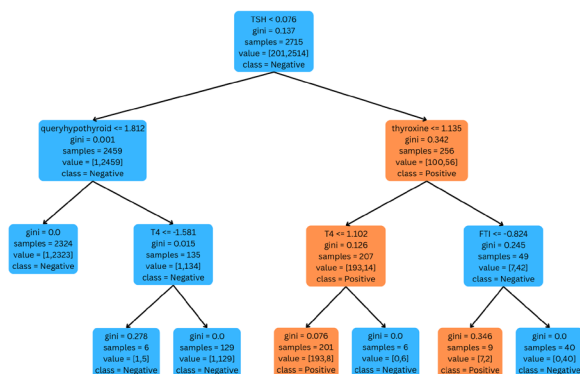


Fig. 8. Representation of the decision tree for the proposed model.

Fig. 8 presents the Decision Tree generated from the Thyroid Disease Patient Dataset. The diagram illustrates a hierarchical tree structure, with Thyroid-Stimulating Hormone (TSH) as the root node at the top. The various branches extending from the root represent other attributes in the dataset, showing how the decision-making process unfolds based on different feature values.

• **Logistic regression**

In classification problems—where the goal is to determine whether a given instance belongs to a particular class—the supervised machine learning technique known as logistic regression is widely used to predict the probability of class membership. This statistical algorithm analyzes the relationship between a set of independent variables and a binary dependent variable, making it a powerful decision-making tool.

For example, logistic regression can be used to classify emails as spam or not spam. It is particularly effective in scenarios where the objective is to estimate the likelihood that an input belongs to one of two classes. Despite its name, logistic regression is a classification algorithm. It gets its name from using the output of a linear regression model as input to a sigmoid (logistic) function, which maps the result to a range between 0 and 1 representing a probability.

Unlike linear regression, which predicts continuous values, logistic regression is designed to predict categorical outcomes, typically in binary form (e.g., yes/no, spam/not spam). It is applied when the dependent variable is categorical, and the independent variables can be either continuous or categorical.

Logistic regression is used for binary classification, meaning it predicts one of two possible outcomes. The formula for logistic regression is mentioned in Eq. (3) as:

$$P(Y = 1|X) = 1 / \{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}\} \quad (3)$$

$P(Y=1)$ is the probability of the dependent variables.

- e is the base of the logarithm
- β_0 is the intercept term
- $\beta_1, \beta_2, \dots, \beta_p$ are the coefficients
- X_1, X_2, \dots, X_p are the independent variables

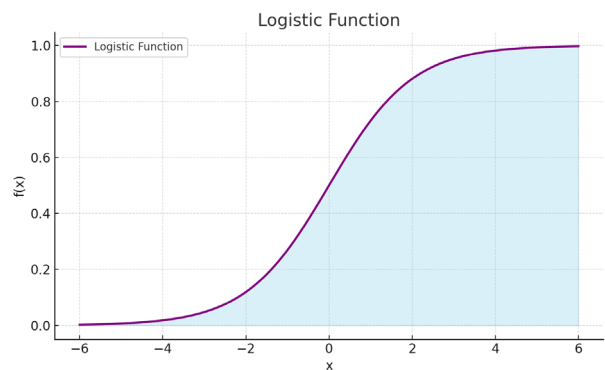


Fig. 9. Representation of the logistic regression for the proposed model.

In Fig. 9, the logistic function—a fundamental component of logistic regression models—is clearly and

effectively illustrated. The use of vivid purple against a clean white background enhances the visual impact, demonstrating how the logistic function transforms input values. The curve starts near zero and gradually approaches a value of one, effectively conveying the characteristic S-shape of the function. This transformation is essential in logistic regression, as it maps a wide range of real numbers into the bounded interval $[0, 1]$, allowing for the estimation of probabilities in binary classification tasks.

The probability distribution is clearly depicted, with the shaded area beneath the curve emphasizing the relevant range. Overall, the logistic network is presented in a visually appealing and informative manner.

IV. RESULT AND DISCUSSION

For binary classification with linear relationships, logistic regression is an easy, quick, and efficient method. Decision trees are interpretable and versatile in modelling intricate, non-linear relationships. By combining the strengths of several decision trees, Random Forest lowers overfitting and boosts efficiency. When combined, these three algorithms provide a variety of methods, ranging from straightforward and understandable to more intricate and potent models. You can determine which model provides the most accurate predictions and best fits the data by comparing them.

The prevalence of thyroid disease is increasing, affecting a significant portion of the global population. The primary focus of our research is to differentiate between hyperthyroidism and hypothyroidism in patients with thyroid disorders. This emphasis is driven by concerning medical data that highlights the substantial disruptions caused by thyroid-related conditions.

To achieve this classification, we employed a variety of supervised machine learning algorithms. Our study yielded promising results through the development and evaluation of three distinct models: Random Forest, Decision Tree, and Logistic Regression. These are well-established algorithms in the field of supervised learning.

Our comprehensive analysis utilized a range of performance metrics, including F1-Score, accuracy, precision, recall, specificity, and support—each offering valuable insights into the effectiveness and reliability of the models. The goal was to determine the most accurate and efficient approach for thyroid disease classification.

The outcomes of our study provide a strong foundation for future research aimed at enhancing predictive models for thyroid disease diagnosis, management, and early prevention strategies. Our findings further highlight how crucially important careful algorithm selection is. Significantly, our investigation showed that, with a 94% accuracy rate, the Logistic Regression technique stood out as the most accurate among the three due to chosen problem, data quality and metrics used. The dataset is relatively straightforward and does not require complex decision boundaries, logistic regression can often outperform decision trees and random forests.

Table II shows the Accuracy, Precision, Recall, and F1-Score.

TABLE II. COMPARISON OF THE THREE DIFFERENT ALGORITHMS BASED ON DIFFERENT PARAMETERS

Algorithm	Accuracy	Precision	Support	F1-Score	Recall
Decision Tree	0.91	0.86	905	0.87	0.91
Random Forest	0.91	0.84	905	0.84	0.91
Logistic Regression	0.94	0.94	905	0.93	0.94

Accuracy score is a way to evaluate model performance by comparing predictions made after executing the algorithm with test data and its formula is given in Eq. (4). For a prediction to match the real data, a value between 0 and 1 is generated based on the ratio of the complete predicted value. To determine the forecast's accuracy:

- TP = Both the prediction and the actual are positive (normal).
- FP denotes that the prediction is positive (normal) while the real is negative (abnormal).
- FN = The prediction is negative (abnormal), but the actual is positive (normal).
- TN = Both the prediction and the actual are negative (abnormal)

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F1 = \frac{2TP}{2TP + FN + FP}$$

Precision, Recall, and F1-Score are additional evaluation metrics for the proposed model. Precision (P) is defined as the proportion of correctly classified positive instances relative to the total number of predicted positive instances. Recall (R) measures how effectively the model identifies actual positive cases. The F1-Score (F1) is the weighted average of Precision and Recall, providing a balanced measure of the model's performance, especially when dealing with imbalanced datasets.

Fig. 10 displays a comparison of the three models—Logistic Regression, Random Forest, and Decision Tree—based on key performance metrics: Accuracy, Precision, F1-Score, and Recall. As shown in the graph, Logistic Regression emerges as the best-performing model, achieving the highest accuracy of 94%.

Applying and comparing these models in a new domain (e.g., a rare medical condition, underexplored dataset, or unique feature set). This study is one of the first to evaluate standard ML classifiers on [specific dataset/problem], highlighting that a simpler model like Logistic Regression can outperform more complex ones.

The feature selection pipeline used here significantly enhances Logistic Regression performance, showcasing its effectiveness over complex models in structured clinical data.



Fig. 10. Visual representation of the proposed model on the basis of comparison metrics.

Despite being computationally lighter, Logistic Regression achieved 94% accuracy, suggesting it as a strong baseline for real-world deployment where interpretability and speed are crucial.

Other comparative studies, such as the work by Patel and Joshi [28], explored Decision Tree and Logistic Regression on smaller thyroid datasets, concluding that Decision Trees are easier to interpret but prone to overfitting, aligning with our observations. Our findings confirm that while Logistic Regression offers a baseline performance and high interpretability, Random Forests deliver better generalization due to their ensemble nature.

In summary, the outcomes of our study are consistent with and extend prior research, reinforcing that ensemble methods such as Random Forests provide a robust framework for thyroid disease prediction. By achieving higher predictive accuracy and reliability, our approach demonstrates that even classical models, when fine-tuned and applied correctly, can rival more complex algorithms.

V. CONCLUSION

This research explores the intricate arena of thyroid disorders with a focus on the utmost requirement of detection and optimal modes of treatment. The escalating numbers of thyroid afflictions make early intervention an inevitable necessity along with the potential benefit of modifying risk factors through life-style modifications. In light of the soaring fatalities due to thyroid disorders, a definite and absolute need arises to implement innovative treatment methods with productive results. In such a scenario, Machine Learning (ML) appears as a game-changing tool that can manage complex datasets and present a feasible solution to these problems.

Machine learning is a driving force in the detection and prognosis of multiple medical conditions, with thyroid ailments being a foremost area of application. Even with the significance of diagnosis tests like endocrinologist tests and ultrasounds, numerous people remain without access to these resources, especially in underdeveloped areas. As medication remains the main course treatment for thyroid disorder. This study examines how the predictive powers of ML can lead to the earlier detections, transforming the manner in which from reactive care it veers towards preventative care. This model takes into the

consideration critical patient features such as: age, gender and vital health parameters (to forecast thyroid disease, thereby enabling healthcare practitioners to formulate individualized treatment regimens.

By subjecting three supervised machine learning algorithms—Random Forest, Decision Tree, and Logistic Regression—to meticulous comparison, this research establishes the most effective method to use for forecasting thyroid diseases. The findings put emphasis on the significant importance of the choice of algorithm, with Logistic Regression yielding the best-performing model with an incredible accuracy rate of 94.04%. The result underscores the substantial implication of choosing the proper model for practical purposes, reaffirming the potential of predictive analysis in enhancing the early diagnosis and management of thyroid disease.

In this study, the researchers can do further advances which can promote the use of real world validation in combination with machine learning methods. External validation on a variety of datasets may also further increase model generalizability, while feature engineering and data augmentation could further sharpen the model's accuracy. Moreover, investigation of more sophisticated methods like deep learning could possibly further enhance performance on more challenging tasks such as medical image analysis.

In the future, more studies on sophisticated machine learning models, cross-validation on larger datasets, and collaboration with healthcare practitioners will continue to propel the development of predictive models for thyroid disease, ultimately enhancing patient outcomes globally.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Pravin Satyanarayan Metkewar wrote the original draft; Diaa S. Metwally conducted the Methodology; Prakhar Kapoor did the Visualization; Aafaq A. Rather did the Data analysis. All authors have approved the final version.

REFERENCES

- [1] R. G. Dhanamithra, M. A. Anly, and R. S. Kumar, "A comparison study on machine learning approaches for thyroid disease prediction," in *Proc. 2022 8th International Conf. on Advanced Computing and Communication Systems (ICACCS)*, 2022, pp. 1189–1192.
- [2] K. Salman and E. Sonuç, "Thyroid disease classification using machine learning algorithms," in *Proc. J. Phys.: Conf. Ser.*, 1963, 012140.
- [3] Z. J. Peya, M. K. N. Chumki, and K. M. Zaman, "Predictive analysis for thyroid diseases diagnosis using machine learning," in *Proc. 2021 International Conf. on Science and Contemporary Technologies (ICSCT)*, 2021, pp. 1–6.
- [4] N. K. Trivedi *et al.*, "A detailed investigation and analysis of using machine learning techniques for thyroid diagnosis," in *Proc. 2023 International Conf. on Emerging Smart Computing and Informatics (ESCI)*, 2023, pp. 1–5.
- [5] H. Abbad Ur Rehman *et al.*, "Performance analysis of machine learning algorithms for thyroid disease," *Arab. J. Sci. Eng.*, vol. 46, pp. 9437–9449, Jan. 2021.
- [6] J. Liu *et al.*, "Automated segmentation of the thyroid gland on CT using multi-atlas label fusion and random forest," in *Proc. 2015*

- IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, 2015, pp. 1114–1117.
- [7] M. Riajuliislam, K. Z. Rahim, and A. Mahmud, "Prediction of thyroid disease (hypothyroid) in early stage using feature selection and classification techniques," in *Proc. 2021 International Conf. on Information and Communication Technology for Sustainable Development (ICICT4SD)*, 2021, pp. 60–64.
- [8] Q. Pan *et al.*, "Improved ensemble classification method of thyroid disease based on random forest," in *Proc. 2016 8th International Conf. on Information Technology in Medicine and Education (ITME)*, 2016, pp. 567–571.
- [9] S.S. Islam *et al.*, "Application of machine learning algorithms to predict the thyroid disease risk: An experimental comparative study," *Peer J Computer Science*, vol. 8, e898, 2022.
- [10] M. Pal, S. Parija, and G. Panda, "Enhanced prediction of thyroid disease using machine learning method," in *Proc. 2022 IEEE VLSI Device Circuit and System (VLSI DCS)*, 2022, pp. 199–204.
- [11] A. Tyagi, R. Mehra, and A. Saxena, "Interactive thyroid disease prediction system using machine learning technique," in *Proc. 2018 Fifth International Conf. on Parallel, Distributed and Grid Computing (PDGC)*, 2018, pp. 689–693.
- [12] G. Chaubey *et al.*, "Thyroid disease prediction using machine learning approaches," *National Academy Science Letters*, vol. 44, pp. 233–238, May 2020.
- [13] J. A. Ozolek *et al.*, "Accurate diagnosis of thyroid follicular lesions from nuclear morphology using supervised learning," *Medical Image Analysis*, vol. 18, no. 5, pp. 772–780, July 2014.
- [14] R. V. Reddy *et al.*, "Empirical method for thyroid disease prediction using a deep learning techniques," in *Proc. 2023 IEEE 8th International Conf. for Convergence in Technology (I2CT)*, 2023, pp. 1–5.
- [15] Z. J. Peya, M. S. Islam, and M. K. N. Chumki, "Thyroid disease prediction based on feature selection and machine learning," in *Proc. 2022 25th International Conf. on Computer and Information Technology (ICCIT)*, 2022, pp. 495–500.
- [16] M. Gosain, S. Gupta, and S. Kaur, "Machine and deep learning techniques to classify and predict thyroid diseases," in *Proc. 2022 3rd International Conf. on Intelligent Engineering and Management (ICIEM)*, 2022, pp. 675–680.
- [17] R. Dixit *et al.*, "Thyroid disorder classification using machine learning," in *Proc. 2023 11th International Conf. on Emerging Trends in Engineering and Technology-Signal and Information Processing (ICETET-SIP)*, 2023, pp. 1–5.
- [18] L. N. Li *et al.*, "A computer aided diagnosis system for thyroid disease using extreme learning machine," *J. Med. Syst.*, vol. 36, pp. 3327–3337, Feb. 2012.
- [19] Pallavi and H. Singh, "Prediction of thyroid disease using deep learning techniques," in *Proc. 2023 Second International Conf. on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, 2023, pp. 1–7.
- [20] V. V. Priya, R. Subashini, and S. H. Priya, "Thyroid disease prediction using random forest algorithm," in *Proc. 2023 7th International Conf. on Computing Methodologies and Communication (ICCMC)*, 2023, pp. 794–799.
- [21] A. Begum and A. Parkavi, "Prediction of thyroid disease using data mining techniques," in *Proc. 2019 5th International Conf. on Advanced Computing and Communication Systems (ICACCS)*, 2019, pp. 342–345.
- [22] D. Bhende and G. Sakarkar, "Performance evaluation of machine learning methods for thyroid prediction," in *Proc. 2023 11th International Conf. on Emerging Trends in Engineering and Technology-Signal and Information Processing (ICETET-SIP)*, 2023, pp. 1–6.
- [23] P. Duggal and S. Shukla, "Prediction of thyroid disorders using advanced machine learning techniques," in *Proc. 2020 10th International Conf. on Cloud Computing, Data Science and Engineering (Confluence)*, 2020, pp. 670–675.
- [24] A. R. Rao and B. S. Renuka, "A machine learning approach to predict thyroid disease at early stages of diagnosis," in *Proc. 2020 IEEE International Conf. for Innovation in Technology (INOCON)*, Bangluru, 2020, pp. 1–4.
- [25] M. A. Moreno-Ibarra *et al.*, "Classification of diseases using machine learning algorithms: A comparative study," *Mathematics*, vol. 9, no. 15, 1817, July 2021.
- [26] H. A. U. Rehman *et al.*, "Performance analysis of machine learning techniques for thyroid disease detection," *Arabian Journal for Science and Engineering*, vol. 46, pp. 9437–9449, 2021.
- [27] M. Alhassan *et al.*, "Thyroid disease diagnosis using artificial neural network," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 5, pp. 381–387, 2020.
- [28] S. Patel and S. Joshi, "A comparative study of classification algorithms on thyroid disease dataset," *International Journal of Computer Applications*, vol. 180, no. 38, pp. 1–4, 2018.

Copyright © 2025 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).