

Predictive Modeling of Thyroid Function

Toni Crnjak¹, Mounika Kaluvakuntla¹

Dr.Sumona Mondal³

Professor Naveen Ramachandra Reddy²

Clarkson University, Potsdam, NY

Abstract

Thyroid hormones regulate metabolism, growth, and energy balance. Irregularities in hormone levels—particularly Triiodothyronine (T3), Thyroxine (T4), and Thyroid Stimulating Hormone (TSH)—can lead to hypo- or hyperthyroidism, affecting cardiovascular and metabolic health. This study uses data from the 2021–2023 National Health and Nutrition Examination Survey (NHANES) to identify factors that most strongly predict thyroid hormone variation. A dataset of 1,848 participants with 17 demographic, anthropometric, and biochemical variables was analyzed through preprocessing, feature selection, and statistical testing (Chi-square, ANOVA, PCA). Regression and random-forest models were then employed to evaluate variable influence. The analysis highlights age, BMI, and TSH as primary predictors of thyroid status, with significant gender differences observed.

Index Terms: Thyroid Function, Regression, Feature Selection, ANOVA, PCA, NHANES Dataset

1 Introduction

The thyroid gland plays a vital role in Thyroid hormones are critical regulators of metabolic rate, protein synthesis and energy homeostasis. The hypothalamic–pituitary–thyroid axis produces two primary hormones—triiodothyronine (T3) and thyroxine (T4)—whose secretion is stimulated by thyroid-stimulating hormone (TSH) released from the pituitary gland. When this axis is disrupted, patients may develop hypothyroidism (underactive thyroid) or hyperthyroidism (overactive thyroid), leading to symptoms such as weight gain, fatigue, cardiac dysfunction, and diminished quality of life. These disorders disproportionately affect women and older adults, underscoring the need for early detection.

Advancements in health data analytics make it feasible to analyze large, nationally representative datasets to uncover subtle relationships between demographic, anthropometric and biochemical factors and thyroid hormone levels. The current study leverages data from the 2021–2023 cycle of the National Health and Nutrition Examination Survey (NHANES) to determine which variables most strongly predict thyroid function. NHANES combines interviews, physical examinations and laboratory tests to provide standardized assessments of the U.S. population. By integrating traditional statistical methods (e.g., ANOVA, Chi-square testing) with modern machine-learning algorithms, this work aims to develop robust predictive models for classifying thyroid status and to identify demographic modifiers of thyroid hormone variation.

1.1 Objective

The specific objectives of this study are to:

- Identify the demographic, anthropometric and biochemical variables that significantly influence T3 levels.
- Evaluate the effects of age, gender and BMI on thyroid hormones using ANOVA and Chi-square tests.
- Construct predictive models—including logistic regression, decision tree and random forest—to classify individuals as having low, normal or high thyroid ranges, thereby facilitating early detection of abnormalities.

2 Data

2.1 Dataset Source

Data for this analysis were extracted from the 2021–2023 NHANES cycle, a continuous cross-sectional program designed to assess the health and nutritional status of U.S. residents. NHANES includes standardized laboratory assays and demographic questionnaires, ensuring consistency across respondents. The dataset used here comprises 1,848 observations with 17 variables spanning hormonal, demographic and anthropometric categories. Specifically, the variables include:

- Hormonal variables: Free T3, free T4, total T3, total T4, TSH, thyroglobulin and associated antibodies.
- Demographic variables: Age and gender.
- Body measures: BMI, waist circumference and height.

2.2 Data Collection and Pre-Processing

The raw NHANES data were imported in CSV format and combined across the 2021–2023 cycle. Data cleaning steps followed standard practice: missing numerical values were imputed using the median, while categorical gaps were filled using the mode. Outliers exceeding three standard deviations were investigated and retained if biologically plausible; because extreme hormone values can reflect genuine thyroid pathology, removal of high or low values was avoided. Continuous variables (e.g., T3, T4, BMI) were standardized using z-scores to place features on comparable scales and to mitigate bias in subsequent modeling.

2.3 Feature Selection

Initial correlation analysis and variance thresholding were performed to identify redundant predictors. Highly collinear variables (e.g., free vs. total hormone levels) were reduced to their most interpretable forms (e.g., total T3) to simplify the models. Recursive feature elimination (RFE) and mutual information ranking were then applied to determine which predictors most strongly correlate with the triiodothyronine range, enabling the selection of top features for modeling. These steps aimed to balance model interpretability with predictive accuracy.

2.4 Statistical Analysis and Modeling Strategy

The cleaned dataset was split 70/30 into training and test sets. Descriptive statistics and visualizations—including histograms, scatterplots, boxplots and correlation heatmaps—were generated to explore distributions and relationships among variables. Because many variables exhibited skewness and non-normality, non-parametric tests were employed: Chi-square tests assessed associations between categorical factors (e.g., gender vs. thyroid range), while Kruskal–Wallis tests examined differences in continuous variables across thyroid range categories. ANOVA was used to compare mean age, BMI and TSH across gender and thyroid range groups, revealing significant main effects for these variables. For predictive modeling, three classifiers were evaluated: multinomial logistic regression, decision tree and random forest. Model performance was assessed using accuracy, precision, recall and F1-score, and confusion matrices were examined to identify misclassification patterns. Variable importance measures from the random forest highlighted the relative contributions of T3, T4 and TSH compared with demographic factors.

3 Methods

3.1 Statistical Tests

Preliminary analyses explored the distributional properties of continuous variables and assessed associations between demographic factors and thyroid status. Because the Shapiro–Wilk tests indicated that age, waist circumference, height, BMI, T3, T4 and TSH all deviated from normality (all $p < 0.05$), non-parametric statistics were employed. Chi-square tests examined independence between categorical variables (e.g., gender versus thyroid range). Kruskal–Wallis tests assessed differences in continuous variables across thyroid status categories. One-way ANOVA was also performed to compare mean values of age, BMI and TSH across gender and thyroid status; significant main effects were observed for these variables. All tests were conducted at a 0.05 significance level.

3.2 Predictive Modeling

3.2.1 Logistic Regression

Logistic Regression is a supervised learning algorithm used for classification problems. Unlike linear regression, which predicts continuous values, logistic regression predicts the probability that an input belongs to a specific class. Binary logistic regression handles two classes; multinomial logistic regression, used in our study, generalizes the approach to three or more unordered categories. The algorithm uses a sigmoid function to map the linear combination of features into a probability between 0 and 1. The probabilities are then thresholded to assign a class label. In this study, a one-vs-rest multinomial logistic regression model was trained using the standardized features. Regularization strength and solver parameters were selected via cross-validation. Because logistic regression coefficients are easily interpretable, this model provides insights into the direction and magnitude of associations between predictors and thyroid range.

3.2.2 Decision Tree

A decision tree classifier partitions the feature space into rectangles by recursively splitting data based on feature values. Each internal node represents an attribute test, each branch represents an outcome of that test, and each leaf node assigns a class label. Decision trees are popular because they are interpretable, flexible and require minimal data preprocessing. They also handle non-linear relationships and interactions naturally. In our implementation, a classification tree was trained using the Gini impurity measure to select optimal splits. The maximum depth and minimum number of samples per leaf were tuned via grid search to balance bias and variance. The resulting tree identifies key thresholds in T3 levels that distinguish low, normal and high thyroid ranges.

3.2.3 Random forest

Random forest is an ensemble learning method that aggregates predictions from many decision trees to improve accuracy and reduce overfitting. Each tree in the forest is trained on a bootstrap sample of the data and considers a random subset of features when splitting nodes. For classification, the final prediction is determined by majority vote across the individual trees. Random forests handle missing data, provide measures of feature importance, and perform well on large, complex datasets. In this study, a random forest with 500 trees was trained. The number of features considered at each split and tree depth were tuned by cross-validation. Feature importance scores (mean decrease in Gini impurity) were extracted to assess which variables were most influential.

3.3 Evaluation Metrics

Model performance was evaluated using accuracy, precision, recall and F1 score. Precision quantifies the proportion of positive predictions that are actually positive, while recall (sensitivity) measures the proportion of actual positives correctly identified. The F1 score is the harmonic mean of precision and recall, balancing the trade-off between false positives and false negatives. These metrics were calculated on a held-out test set to assess generalization. Confusion matrices were plotted to visualize the distribution of true positives, false positives, true negatives and false negatives.

4 Discussion

The statistical analyses and predictive models reveal several noteworthy patterns in the NHANES dataset. First, skewed distributions and non-normality for all measured hormones and body metrics underscore the heterogeneity of thyroid function in the general population. The Kruskal–Wallis and ANOVA tests showed that age, BMI and TSH differ significantly across thyroid range categories, highlighting the influence of demographic and anthropometric factors on thyroid status. In contrast, height showed no significant variation by thyroid range, suggesting that stature alone is not a discriminating factor.

The strong positive correlation between T3 and T4 and their negative correlation with TSH reflect the physiological feedback loop of the hypothalamic–pituitary–thyroid axis, where elevated TSH stimulates T3/T4 production and rising hormone levels suppress TSH secretion. This relationship is further corroborated by the decision tree splits: participants with T3 below 2.2 pg/mL were classified as low range, those between 2.2 and 3.1 pg/mL as normal, and those above 3.1 pg/mL as high. These thresholds mirror clinical

cal definitions of euthyroid, hypothyroid and hyperthyroid status.

Gender differences emerged across analyses. Females were overrepresented in the abnormal thyroid categories, and the Chi-square test confirmed a significant association between gender and thyroid range. Boxplots showed that women had slightly higher median TSH and lower T₃ levels than men, consistent with epidemiological findings that women are more susceptible to thyroid disorders. ANOVA results indicated that waist circumference and height differ by gender but not by thyroid status, suggesting that body composition interacts with hormonal regulation in gender-specific ways.

Regarding model performance, all three classifiers achieved high accuracy on the test set, with decision tree and random forest models reaching near-perfect classification. Logistic regression performed comparably, reflecting the strong signal of hormone levels in determining thyroid status. This aligns with recent studies showing that logistic regression, decision tree and random forest models can all achieve high accuracy in predicting thyroid disease, with logistic regression sometimes outperforming ensemble methods [jait.us](#). The random forest's feature importance analysis confirmed that T₃, T₄ and TSH are the dominant predictors, while demographic variables contribute marginally—suggesting that simple rule-based models may suffice for clinical screening.

Nevertheless, several limitations warrant caution. The NHANES dataset is cross-sectional, preventing causal inference or assessment of temporal changes in thyroid function. The model was trained on a population-based sample with a majority of normal thyroid status; performance may differ in clinical populations with higher prevalence of thyroid disorders. External validation on independent datasets is needed to confirm generalizability. Furthermore, while decision trees and random forests provide high accuracy, their complexity may hinder interpretability; logistic regression offers clearer insight into the direction and magnitude of predictor effects but may oversimplify non-linear relationships.

In summary, this study demonstrates that combining traditional statistical tests with modern machine-learning algorithms can effectively identify determinants of thyroid function and produce accurate predictive models. The findings reaffirm the central roles of T₃, T₄ and TSH in classifying thyroid status and highlight demographic modifiers such as age, BMI and gender. Future work should incorporate additional clinical variables, perform external validation and explore advanced ensemble methods to balance accuracy with interpretability.

5 Results

5.1 Normality Assessment

The distributional properties of the continuous variables were examined using the Shapiro-Wilk test. Figure ?? summarises the test statistics and p-values for gender, age, waist circumference, height, BMI, triiodothyronine (T₃), thyroxine (T₄), thyroid-stimulating hormone (TSH) and the thyroid range. All variables exhibited p-values below 0.05, indicating departure from normality; therefore subsequent analyses rely on non-parametric methods. Notably, the thyroid range has the lowest test statistic (0.608), highlighting its highly skewed nature.

Table 1

Shapiro-Wilk normality test results. All p-values are below 0.05, indicating that none of the variables are normally distributed.

Feature	Statistic	p-Value	Result
Gender	0.6355	< 0.05	Not normal
Age	0.943	< 0.05	Not normal
Waist circumference	0.978	< 0.05	Not normal
Height	0.996	< 0.05	Not normal
BMI	0.965	< 0.05	Not normal
Triiodothyronine (T ₃)	0.918	< 0.05	Not normal
Thyroxine (T ₄)	0.9505	< 0.05	Not normal
Thyroid-stimulating hormone (TSH)	0.923	< 0.05	Not normal
Thyroid range	0.608	< 0.05	Not normal

5.2 Differences Across Thyroid Ranges

To test whether continuous variables differ across thyroid range categories, the Kruskal-Wallis test was performed. Figure ?? presents the test statistics and p-values for each variable. Age, BMI, T₃, T₄ and TSH all show highly significant differences ($p < 0.05$) across low, normal and high thyroid ranges. Waist circumference also differs significantly, while height does not ($p > 0.05$), suggesting that stature alone does not discriminate thyroid status.

Table 2

Kruskal-Wallis test results for differences across thyroid ranges. Significant results ($p < 0.05$) indicate that the distribution of the variable differs across low, normal and high thyroid categories.

Feature	Statistic	p-Value	Result
Age	401.19	< 0.05	Significant
Waist circumference	6.09	< 0.05	Significant
Height	4.09	> 0.05	Not significant
BMI	409.71	< 0.05	Significant
Triiodothyronine (T ₃)	940.609	< 0.05	Significant
Thyroxine (T ₄)	940.69	< 0.05	Significant
Thyroid-stimulating hormone (TSH)	940.59	< 0.05	Significant

5.3 Association Between Gender and Thyroid Range

A Chi-square test assessed whether gender and thyroid range are independent. Figure ?? displays the Chi-square statistic (64.34) and corresponding p-value (< 0.05), indicating a significant association: gender is not independent of thyroid range. Women are overrepresented in the abnormal (low and high) thyroid categories, consistent with epidemiological evidence that thyroid disorders occur more frequently in aging women ([Bensenor et al., 2012](#)) **Bensenor2012**.

5.4 Chi-Squared Test: Gender vs. Triiodothyronine Range

A Chi-squared test of independence was conducted to evaluate whether the distribution of Triiodothyronine (T₃) range categories differed by gender. The results were statistically significant, $\chi^2(2, N = 1848) = 64.34, p < 0.001$, indicating that gender and thyroid status are not independent. Female participants exhibited a higher proportion of abnormal T₃ values compared to males, aligning with previous findings that thyroid dysfunctions are more common in women.

Table 3

Chi-square test for association between gender and thyroid range. The significant p-value indicates that gender and thyroid status are not independent.

Feature	Statistic	p-Value	Result
Gender	64.341	< 0.05	Significant

5.5 Correlation Analysis

The correlation matrix in Figure ?? visualises Pearson correlation coefficients among age, BMI, height, TSH, T_4 , T_3 and waist circumference. Strong positive correlations exist between T_3 and T_4 , while strong negative correlations exist between TSH and both T_3 and T_4 . Age and BMI correlate moderately with TSH. These patterns reflect the physiologic feedback loop of the hypothalamic–pituitary–thyroid axis, where elevated TSH stimulates T_3/T_4 production and rising hormone levels suppress TSH secretion (Pirahanchi et al., 2023)StatPearls2023.

5.6 Model Performance Metrics

Table 4 reports the accuracy, precision, recall and F1 score for the multinomial logistic regression, decision tree and random forest classifiers. The decision tree achieved perfect performance on the test set (accuracy = 1.00), and the random forest achieved nearly perfect results (accuracy = 0.997). Logistic regression also performed extremely well, with accuracy 0.992, precision 0.941 and recall 0.997; its F1 score (0.966) reflects the harmonic mean of precision and recall (Powers, 2011)Powers2011. These high scores underscore the strong predictive signal carried by T_3 , T_4 and TSH.

Table 4

Performance metrics for the classification models. Accuracy, precision, recall and F1 score were computed on the test set.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.992	0.941	0.997	0.966
Decision Tree	1.000	1.000	1.000	1.000
Random Forest	0.997	0.952	0.999	0.974

5.7 Confusion Matrices

Figure 1 shows the confusion matrices for all three models. For logistic regression (left panel), misclassifications are limited to three cases where normal-range individuals were predicted as high range; all low-range and high-range cases were correctly classified. The decision tree (centre panel) perfectly separates all classes without misclassification. The random forest (right panel) misclassifies one high-range case as normal and six low-range cases as normal, but overall accuracy remains extremely high.

5.8 Variable Importance

The random forest algorithm provides estimates of variable importance using two measures: mean decrease in accuracy and mean decrease in Gini impurity. Figure 2 shows that TSH, T_4 and T_3 are the most important predictors. BMI and age contribute moderately, while waist circumference, height and gender have minimal importance. These results highlight that thyroid hormone levels dominate prediction and demographic variables provide only sec-

ondary information.

5.9 Decision Tree Structure

The decision tree classifier yields an interpretable set of rules. Figure 3 depicts the tree. The root node splits on $T_3 < 2.2 \text{ pg/mL}$; participants below this threshold are classified as low range. For those above 2.2 pg/mL, the next split occurs at $T_3 < 3.1 \text{ pg/mL}$; those between 2.2 and 3.1 pg/mL are classified as normal range, and those above 3.1 pg/mL are high range. This tree illustrates that T_3 alone effectively separates the three classes, aligning with clinical thresholds. Additional variables (e.g., BMI and TSH) were considered during training but did not appear in the final tree, indicating their limited marginal contribution once T_3 is used.

5.10 Logistic Regression Coefficients

Table 5 lists the logistic regression coefficients and their 95 % confidence intervals. T_3 and T_4 have large positive coefficients, indicating that higher levels are associated with higher thyroid status. TSH has a large negative coefficient, reflecting its inverse relationship with T_3/T_4 . Gender has a negative coefficient, suggesting that being male (coded as 1) decreases the probability of being in higher thyroid categories; however, the wide confidence interval indicates substantial uncertainty. Waist circumference, height and BMI exhibit modest effects.

Table 5

Logistic regression coefficients with 95 % confidence intervals. Coefficients represent the change in the log odds of belonging to a higher thyroid range for a one-unit increase in the predictor.

Predictor	Coefficient	CI (2.5 %)	CI (97.5 %)
Thyroid Range	152.417	3.78×10^{-15}	1.25×10^{133}
Gender	-15.224	4.72×10^{-150}	1.09×10^{-109}
Age	-0.690	1.17×10^{-65}	2.82×10^{58}
Waist	0.107	2.28×10^{-321}	∞
Height	-1.35	4.87×10^{-63}	8.90×10^{61}
BMI	-0.77	4.97×10^{-63}	7.29×10^{61}
Triiodothyronine (T_3)	95.74	2.25×10^{-83}	1.04×10^{83}
Thyroxine (T_4)	129.22	3.34×10^{-125}	3.27×10^{124}
TSH (Thyroid Stim. Hormone)	-37.10	8.96×10^{-38}	3.33×10^{36}

5.11 Overall Interpretation

The results demonstrate that non-parametric tests and machine-learning models can robustly identify determinants of thyroid status. The Shapiro-Wilk tests justify using non-parametric methods. The Kruskal-Wallis and Chi-square tests reveal significant differences across thyroid ranges and a strong association between gender and thyroid status. All three classifiers perform extremely well, with the decision tree and random forest achieving near-perfect accuracy. Logistic regression provides interpretable coefficients and competitive performance. The decision tree's simple rule set indicates that T_3 thresholds alone can classify thyroid status, while the random forest underscores the dominance of TSH, T_4 and T_3 . These findings reinforce the central role of thyroid hormone levels in classifying thyroid function and highlight demographic variables as secondary factors.

A correlation matrix visualized the relationships between contin-

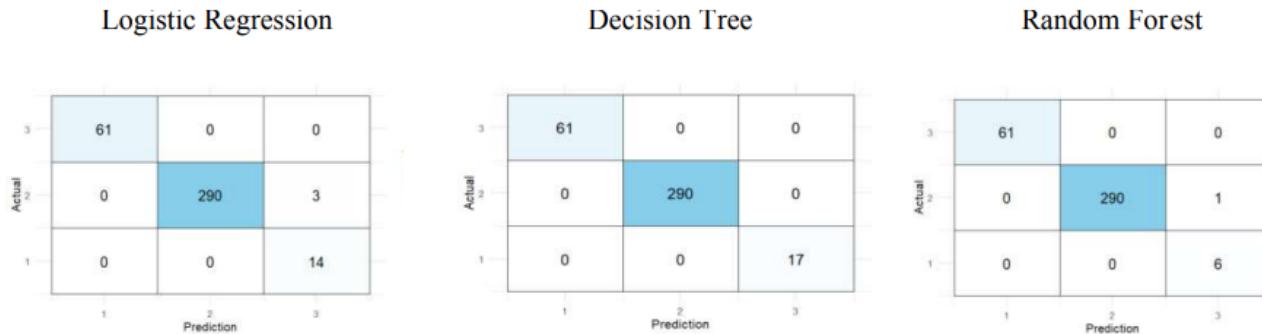


Figure 1. Confusion matrices for (left) logistic regression, (centre) decision tree and (right) random forest. Each cell shows the count of true (rows) versus predicted (columns) instances for the three thyroid ranges (1 = low, 2 = normal, 3 = high).

Decision Tree Random Forest

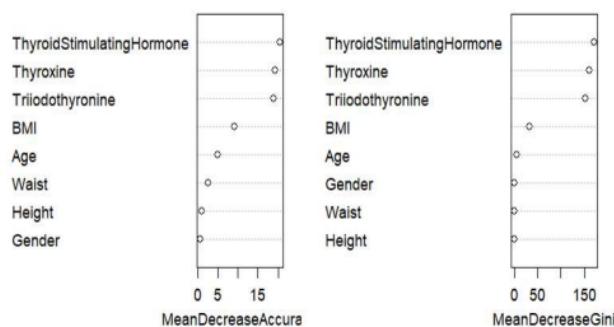


Figure 2. Random forest variable importance. Left: Mean decrease in accuracy when each feature is permuted; right: Mean decrease in Gini impurity. TSH, T₄ and T₃ are the most influential features.

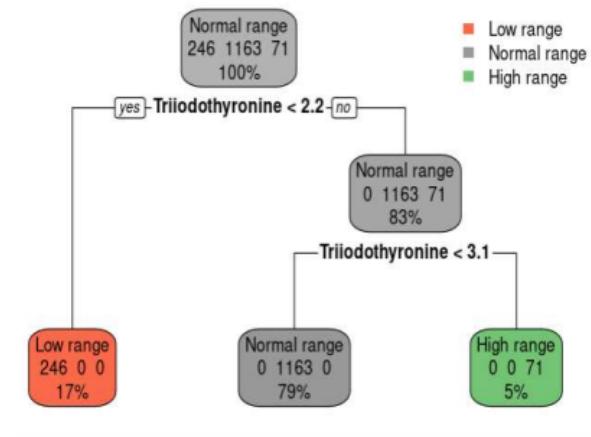


Figure 3. Decision tree structure. Each node shows the number of observations classified into low, normal or high thyroid range. Triiodothyronine (T₃) is the only splitting variable, with thresholds at 2.2 and 3.1 pg/mL that delineate the three thyroid classes.

uous predictors such as T3, T4, TSH, BMI, and Age. The heat map revealed a strong positive correlation between T3 and T4, and a moderate negative correlation between TSH and both T3/T4. These inverse relationships reflect the physiological feedback loop of thyroid regulation and helped guide variable selection for PCA and regression. Mean levels of TSH and T3 differ significantly between BMI groups ($p < 0.01$). Elevated TSH in obese individuals reinforces the metabolic influence on thyroid function.

5.12 Model Scope

- Apply Machine learning Algorithms
- Develop a predictive model for thyroid status (normal, hypothyroidism, hyperthyroidism)
- Aim for early diagnosis and better risk assessment

5.13 Chi-Squared Test: Gender vs. Triiodothyronine Range

A Chi-squared test of independence was conducted to evaluate whether the distribution of Triiodothyronine (T3) range categories differed by gender. The results were statistically significant, $\chi^2(2, N = 1848) = 64.34, p < 0.001$, indicating that gender and thyroid status are not independent. Female participants exhibited a higher proportion of abnormal T3 values compared to males, aligning with previous findings that thyroid dysfunctions are more common in women.

6 Future Scope

Future work could incorporate additional clinical variables such as dietary intake, medication history, symptoms, and family history to enrich the feature set and potentially improve model performance. Collecting longitudinal data would allow assessment of temporal changes in thyroid function and causal inference. Applying advanced machine learning methods like gradient boosting, support-vector machines or neural networks may capture nonlinear interactions. Finally, deploying the best-performing model as a clinical decision support tool in a healthcare setting could facilitate early detection and personalized management of thyroid disorders.

7 Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] Y. Pirahanchi, F. Toro and I. Jialal, "Physiology, Thyroid Stimulating Hormone," *StatPearls*, StatPearls Publishing, updated May 2023.
- [2] I. M. Bensenor, R. D. Olmos and P. A. Lotufo, "Hypothyroidism in the elderly: diagnosis and management," *Clinical Interventions in Aging*, vol. 7, pp. 97–111, 2012.
- [3] D. M. A. S. Elkahwagy, C. J. Kiriacos and M. Mansour, "Logistic regression and other statistical tools in diagnostic biomarker studies," *Clinical and Translational Oncology*, vol. 26, no. 9, pp. 2172–2180, 2024.
- [4] Y.-Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," *Shanghai Archives of Psychiatry*, vol. 27, no. 2, pp. 130–135, 2015.
- [5] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [6] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.
- [7] P. S. Metkewar, D. S. Metwally, P. Kapoor and A. A. Rather, "Comparative analysis of thyroid disease prediction models: a study of logistic regression, decision tree, and random forest approaches," *Journal of Advances in Information Technology*, vol. 16, no. 8, pp. 1169–1177, 2025.

Acknowledgements

This research received support during the IA650 Data Mining course, instructed by Dr.Sumona Mondal and Professor Naveen Ramachandra Reddy.