FUDMA Journal of Engineering and Technology
ISSN: 3092-9385
https://fjet.fudutsinma.edu.ng/index.php/fjet/
Volume 2, Issue 1, 2026, 640-653

Machine Learning Based Feature Selection for Early Detection of Thyroid Disorders in Nigeria

Peter S. IDOKO[1*], Iyinoluwa T. IDOWU[2], Emmanuel O. AYODELE[3], Temitope F. SHOLANKE[4], Peter A. IDOWU[5]

[1]Department of Computer Science, Federal Polytechnic Orogun, Delta State, Nigeria
[2]Department of Nursing Science, Elizade University, Ilara-Mokin, Ondo State, Nigeria
[3]Computer Science Program, Bowen University, Iwo, Osun State, Nigeria
[4,5]Department of Computer Science and Cybersecurity, Obafemi Awolowo University, Ile-Ife, Nigeria

[1]idoko.peter@fepo.edu.ng, [2]idowuiyinoluwa056@gmail.com, [3*]emmanuel.ayodele@bowen.edu.ng, [4]tsholanke@oauife.edu.ng, [5]paidowu@oauife.edu.ng

Abstract
Disorders of the thyroid are regarded as one of the main concerns related to global public health issues. They cause considerable harm in underdeveloped countries like Nigeria. This paper attempts to find the most accurate predictors of Nigerian thyroid disease via using the machine learning approach. Finding relevant features for building the models with robust forecasting reliability is a crucial stage of the machine learning process. All redundant variables should be eliminated from the initial data set for the sake of making the process of model training more effective and avoiding possible cases of overfitting. That is why this paper aims at using machine learning methods for selecting those features suitable for predicting the early development of the disease in the Nigerian population. Clinical indicators (TSH, T3, T4, autoantibodies), demographic parameters (sex, age, body mass index), ultrasound characteristics, and environmental variables (exposure to goitrogens and iodine content) are taken into account. Both a filtering approach and the usage of Random Forest algorithm are utilized to select the best features. As shown by results, Random Forest and Gradient Boosting performed equally well, while Random Forest has slightly better predictive power. Using the entire set of features, Random Forest reached the accuracy of 0.9978, a precision of 0.9986, a recall of 0.9971, F1-score of 0.9978, and an ROC-AUC equal to 0.9999. Gradient Boosting demonstrated the same performance: accuracy = 0.9971, ROC-AUC = 0.9999.

Keywords: *Gradient Boosting, Machine Learning, Random Forest, Selective Features, Thyroid Disorders.*

1.0 Introduction

Thyroid disorders represent a major health issue on a global scale, especially in the less developed nations such as Nigeria, where delayed diagnosis and poor health infrastructure are prevalent drivers of poor patient outcomes. Located in the neck region, the thyroid is an important endocrine organ involved in growth and metabolism by producing thyroxine (T4) and triiodothyronine (T3) in response to the stimulation of the pituitary gland by the thyroid-stimulating hormone (TSH) [1]. Imbalances in these complex physiological mechanisms lead to various forms of thyroid disorders such as goiter, hyperthyroidism, hypothyroidism, and thyroid cancer. Some common clinical manifestations of these disorders include fatigue, weight gain/loss, cardiac issues, and malignant thyroid disease [2]. In Nigeria, the prevalence of thyroid disorders among the general population ranges between 15% to 20%, although some subpopulations exhibit high incidences attributable to geographical factors [3]. High levels of disease incidence combined with low public knowledge about thyroid disorders and lack of advanced testing methods necessitate the development of novel approaches for screening thyroid-related diseases.

Due to its dense population, Nigeria faces unique challenges in handling diseases involving the thyroid gland. There is a shortage of endocrinologist specialists in the country's healthcare system. Moreover, the lack of access to thyroid function tests and ultrasound scans (TSH, T3, T4, anti-Tg, anti-TPO) hinders early diagnosis. In addition, economic issues, such as the widespread problem of poverty and poor medical facilities in rural areas, make early diagnosis difficult. Indeed, fine needle aspiration cytology (FNAC), an important test for the detection of thyroid nodule, is mostly available in large urban hospitals [2]. The problems described above result in delayed clinical diagnosis. At the same time, Nigeria has been recording a growing number of thyroid cancer cases. Today, papillary thyroid carcinoma makes up almost 53% of newly diagnosed patients [3]. Dietary habits and environmental factors are instrumental in determining the epidemiology of diseases associated with thyroid. The most common causes of thyroid dysfunction are hypothyroidism and goiter due to iodine deficiency, that are traceable in some parts of Nigeria, particularly in the north, where there is a deficiency of iodine in drinking water and soil [4]. Despite the existent of programs encouraging people to use iodized salt, their effectiveness appears to be insufficient. Findings shows that only 60% of household use properly iodized salt in northern Nigeria [5].

Received: 12/03/2026; Revised: 18/03/2026; Accepted: 19/03/2026; Published: 09/05/2026

Another problem affecting the state of health of residents of this region is their heavy consumption of foods containing goitrogens (e.g., cassava). These foods interfere with hormone production and the absorption of iodine [3]. It has been established that goiters appear due to the interaction between cassava-based foods and the thyroid gland. Cassava has cyanogenic glycosides, and upon digestion, they form thiocyanate compounds, which inhibit iodine uptake in thyroid [4]. Furthermore, additional environmental variables lead to thyroid dysfunction. These include genetic susceptibility to autoimmunity and exposure to radiation.

Currently, machine learning (ML) acts as a revolutionary tool in the field of medicine, as ML may be used to tackle diagnostic challenges in developing countries. Neural Networks, SVM, and Random Forest algorithms can process complex datasets with the inclusion of environmental, demographic, and medical parameters to accurately assess the chances of disease development [6]. When it comes to thyroid disorders, there are ML algorithms which are successfully implemented worldwide and can diagonise patients with thyroid cancer, hyper- and hypothyroidism using patient demographics, images, and laboratory tests [1]. According to [6], the Neural Network algorithm was able to predict thyroid disease diagnosis using autoantibodies in combination with levels of TSH, T3, and T4, with an impressive F1-score of 0.92. Conversely, the ability of Random Forest algorithm to process high-dimensional data allows including various parameters such as diet information and ultrasound results [2].

The use of ML algorithms to assess thyroid disease risks in Nigeria is quite advantageous because ML-based classifiers are known to process fragmented and diverse data well, which might be the case for Nigerian healthcare. With regional risk factors included in ML classification (such as socioeconomic status, goitrogens, and iodine deficiency), a cost-effective and scalable system of thyroid disorders preliminary identification can be created [5]. This will help to allocate patients who require further diagnosis at hospitals, reducing pressure on medical resources.

In conclusion, this paper aims to examine a ML-based feature selection model to predict the development of thyroid disorders in Nigeria. This research considers a number of parameters, namely environmental factors (goitrogen exposure, iodine concentration), demographic parameters (patient BMI, gender, age), clinical features (autoantibodies, T4, T3, TSH), and ultrasounds.

2.0 Related works

Research by [7] looked into forecasting thyroid diseases using feature selection and machine learning. They eliminated unnecessary attributes from the UCI repository dataset to focus on the variables with the greatest predictive power. The dataset included three main categories: hyperthyroid, hypothyroid, and normal. They processed the data with seven different algorithms. In the end, the Random Forest model outperformed other modern methods, achieving an accuracy of 99.58%. Nevertheless, their investigation failed to address the specific regional risk factors related to Nigeria, which is the main goal of this current study.

An investigation by [8] used deep learning methods to select features for classifying thyroid conditions. They categorized thyroid disorders into three groups: normal, hypothyroidism, and hyperthyroidism, and used these categories for diagnostic support. Their work mainly focused on improving the model and engineering features within a deep learning framework. To increase accuracy, they extracted features using an extra tree classifier along with a random forest algorithm. Their results showed strong performance metrics in assessing thyroid nodules, which could significantly help in clinical diagnosis. The proposed model's effectiveness was further confirmed through K-fold cross-validation and F1-score analysis. Nonetheless, their investigation make no explicit mention of for the best predictive features for general thyroid disease detection, a gap this paper aims to address.

A study by [9] focused on a diagnostic approach using features derived from Random Forest to achieve higher predictive accuracy. Their method successfully identified ten different thyroid conditions with an accuracy of 0.99. Despite these robust results, the research did not conduct a thorough search for the very best predictive features, setting it apart from the objectives of our current analysis.

[10] analyzed how wrapper-based (Recursive Feature Elimination) and filter-based (F-Score) feature selection techniques affect disease categorization and identification. Their study included dimensionality reduction using Principal Component Analysis. They evaluated performance based on three main metrics: specificity, sensitivity, and accuracy. The algorithms were tested on four models: Extreme Learning Machine, Support Vector Machine, Back Propagation Neural Network, and MultiLayer Perceptron. The findings showed that while both recursive elimination and F-Score methods improved diagnostic results, the wrapper-based method was the most effective, achieving a peak accuracy of 98.14% with the ELM classifier. Much like prior investigation, this study omitted empahsis on isolating the best predictive features for thyroid dysfunction.

[11] researched thyroid abnormalities, noting that these issues arise from gland problems that cause metabolic imbalances due to irregular hormone production. An underactive gland leads to hypothyroidism, while an overactive gland causes hyperthyroidism. If not addressed, both situations can become serious health emergencies. Timely detection is crucial to avoid severe complications and maintain a good quality of life through accurate hormonal balance. Their hybrid algorithmic framework identified key features using metrics like recall, precision,

Received: 12/03/2026; Revised: 18/03/2026; Accepted: 19/03/2026; Published: 09/05/2026

F1-score, and accuracy. When tested against a benchmark dataset, their method produced strong results, including an F1-score of 94.83 and an accuracy of 98.91%. This study advanced medical diagnostics by combining nature-inspired optimization models with machine learning to detect thyroid diseases early. The authors suggested merging Simulated Annealing (SA) with the Cuttlefish Optimization Algorithm (CFA) to identify optimal disease features. Using these optimization techniques significantly enhances healthcare diagnostics. However, their main focus was not solely on the isolation of raw features for prediction, which is the main focus of our paper.

[12] introduced a new prediction model that analyzed three different open-source datasets. To eliminate bias, they balanced and standardized the raw data during preprocessing. They then used a cascaded autoencoder-based simple recurrent architecture to extract important spatio-temporal properties. To improve the model's efficiency, they selected optimal features using a newly proposed Opposition Learning-based Red Panda Optimization (OL_RPO) algorithm. The final prediction phase used an Enhanced Transformer Model to generate accurate and reliable forecasts. Their evaluation metrics (NPV, PPV, F-Score, Sensitivity, Specificity, Accuracy, and Error) were 1.9, 98.1, 98.501, 99.01, 99.2, 99, and 0.07689, respectively. Because their work emphasized deep learning over traditional feature selection, it differs significantly from the focus of our research.

3.0 Method

We sourced significant volume of data from a free public site. This data is about patients who have thyroid problems. There are 3,772 patients in the data and 26 different things we know about each patient. Some of this information is just categories, like yes or no, and some is numbers. We took this data and put it into a special format that computers can understand. The first line of the data tells us what each piece of information means, and the rest of the lines tell us about each patient. We want to know if a patient has a thyroid problem or not, so we made that the main thing we're trying to figure out. We used the other 25 pieces of information to help us guess if a patient has a thyroid problem. The data tells us a lot about each patient, like their medical history, what medicines they're taking, and what their hormone levels are. We made a table that shows what each piece of information means and what kind of data it is. This helps us understand the data better.

This repository was evaluated to map out the connections between biochemical indicators, clinical traits, and thyroid anomalies, laying the groundwork for precise ML-based early detection models.

Table 1: Description of identified variables

Variables	Data Type	Description
Age	float64	Patient age in years (numerical)
Sex	float64	Gender of patient (0 = male, 1 = female)
on thyroxine	int64	Currently taking thyroxine (1 = yes, 0 = no)
query on thyroxine	int64	Suspected thyroxine use (1 = yes, 0 = no)
on antithyroid medication	int64	On antithyroid meds (1 = yes, 0 = no)
Ill	int64	ill patient (1 = yes, 0 = no)
pregnant	int64	Pregnant patient (1 = yes, 0 = no)
thyroid surgery	int64	History of thyroid surgery (1 = yes, 0 = no)
I131 treatment	int64	Received I131 treatment (1 = yes, 0 = no)
query hypothyroid	int64	Suspected hypothyroidism (1 = yes, 0 = no)
query hyperthyroid	int64	Suspected hyperthyroidism (1 = yes, 0 = no)
Lithium	int64	On lithium medication (1 = yes, 0 = no)
goitre	int64	Presence of goitre (1 = yes, 0 = no)
tumor	int64	Presence of tumor (1 = yes, 0 = no)
hypopituitary	int64	Hypopituitary condition (1 = yes, 0 = no)
psych	int64	Psychological disorders (1 = yes, 0 = no)
TSH measured	int64	TSH test done (1 = yes, 0 = no)
T3 measured	int64	T3 test done (1 = yes, 0 = no)
TT4 measured	int64	Total T4 test done (1 = yes, 0 = no)
TT4	float64	Total thyroxine level (μg/dL)
Diet	Int64	Balance Diet(1 = yes, 0 =no)
binaryClass	int64	Target variable (1 = thyroid disorders, 0 = normal)

3.1 Method of preprocessing of collected Data

When we're building a reliable machine learning framework, it's really important to verify and clean the data. To do this, we need to go through a process called data cleaning. We used a breakdown of our dataset, shown in Table 1, to guide us through this phase. The goal is to make sure the information we're feeding into our algorithms is correct and consistent. One of the things we had to do was fill in missing values in our dataset. This is called

Received: 12/03/2026; Revised: 18/03/2026; Accepted: 19/03/2026; Published: 09/05/2026

imputation, and it's necessary to circumvent data degradation. If we don't fill in these gaps, we might end up with inaccurate results. We especially needed to fill in missing values for certain clinical markers, like FTI, TT4, and TSH, because these were often left blank when laboratory tests weren't done. Another important step was standardizing our features. We used a tool called StandardScaler from the scikit-learn library to do this. What this does is adjust all our numerical data so that it has a standard deviation of 1 and a mean of 0. This is really important because some of our variables had very different scales, and this can affect how well our classifiers work. Classifiers like Gradient Boost and Random Forest need our data to be standardized so they can make accurate predictions. By doing these steps, we can ensure the integrity and precision of our dataset, which is essential for building a good machine learning framework.

The following utilities were mobilized for the data preparation pipeline:

i. Pandas for structural management and dataset manipulation.
ii. Scikit-learn for splitting the data (using train_test_split), feature transformation, and scaling.
iii. NumPy for executing rapid numerical computations.

This tough data cleanup made sure the information was ready for the next steps, which helped make the final results easier to understand and more accurate. By doing this, we were able to derive more precise insights from our model and make more sense of the data. This is important because it helps us trust the results and make good decisions based on them.

3.2 Method of feature selection

Finding the right set of variables is crucial when creating accurate models to diagnose thyroid disease. To do this, we used a two-part approach, combining filter-based methods with Random Forest algorithms to ensure we got the most important predictors. This way, we could be sure that our models would be as accurate as possible. The filter-oriented techniques helped us narrow down the options, and then the Random Forest algorithms further refined our selection. By using both methods together, we were able to identify the most influential variables that really make a difference in diagnosing thyroid disease.

Filter methods look at variables on their own, without using the main learning algorithm. They use statistical attributes and how intimately they are intertwined with the target class. This approach is simple and can be used in many situations, making it perfect for a first look at large medical datasets where clinical variables have different levels of importance. The filter-based analysis started by looking at correlations to find variables that overlap and could cause problems. Variables with correlation scores over 0.9 were marked for removal. Also, univariate statistical checks were used to see how valuable each feature was on its own. For numeric variables, the F-statistic was used to see how closely they were related to the target disorder. For categorical data, chi-square tests were used to make sure they were independent of the target variable. Information gain was also used as a key metric to measure how much the uncertainty decreased when data was divided based on specific traits. Variables with high information gain were preferred because they were good at telling apart clinical categories. This was backed up by mutual information scoring, which can show both non-linear and linear relationships. The goal is to find the most useful variables and get rid of the ones that don't add much value. By using these filter methods, we can simplify the data and make it easier to analyze, which is especially important when working with large medical datasets. In the end, the filter methods help us choose the best variables for our analysis, which can lead to better results and a deeper understanding of the data. They are a crucial step in the process, allowing us to focus on the most important variables and ignore the ones that don't contribute much to our understanding of the target class.

At the same time, a Random Forest approach was used to create rankings that show how important each feature is. This method is different from simple filter techniques because it takes into account the complex and non-linear relationships that are common in healthcare datasets. The algorithm builds many decision trees using random subsets of features and bootstrap sampling, which helps to prevent overfitting and keeps the predictions accurate. In a Random Forest, the value of each feature is appraised in two ways. First, it looks at how much a feature helps to create homogeneous groups of data; features that do a good job of splitting the data into pure groups get higher rankings. Second, it looks at how much the predictions get worse when a feature is randomly changed, which shows how much that feature contributes to the model's success. To get these rankings, we trained a Random Forest with 100 decision trees on the cleaned dataset. When building each tree, we randomly selected features to use, based on the square root of the total number of features. We also used bootstrap sampling to make sure the importance scores were reliable, by training each tree on a different subset of the data.

Merging the Random Forest and filter methodologies yielded a highly robust evaluation environment. Attributes flagged as crucial by both protocols were pushed to the forefront for final model inclusion. This dual-validation tactic successfully strips away irrelevant noise, ensuring only genuinely predictive factors survive. To harmonize the outcomes, a unified ranking matrix was developed: filter metrics were adjusted to a standard scale,

Received: 12/03/2026; Revised: 18/03/2026; Accepted: 19/03/2026; Published: 09/05/2026

while Random Forest scores underwent z-score transformations. This synthesized hierarchy factored in both predictive weight and statistical relevance, forging a solid base for the final subset.

We tried different combinations of top variables to find the best balance between how fast the computer can process them and how accurate the diagnosis is. This was done through a process of testing and retesting, called cross-validation, to make sure the final set of variables was the best it could be for telling classes apart.

4.0 Results and discussion

4.1 Exploration data analysis of thyroid disorders dataset results

When we looked at the data from 3,772 patient logs about thyroid function, we wanted to see how different factors were connected. We had information about the patients' age, sex, and some clinical markers, like whether they were taking thyroxine. We also had data on hormone levels, such as FTI, TT4, and TSH, as well as details about their diet and a classification of their health as either positive or negative. Our goal was to understand how these clinical metrics behaved, how they related to thyroid disease, and how diet affected things. To make sure our analysis was accurate, we cleaned up the data thoroughly. We made sure the numeric columns were correct, and we limited the binary factors - like diet, health classification, sex, and thyroxine use - to either 0 or 1. We also got rid of any rows with corrupted data. For the remaining missing numeric values, we filled them in with the average value for that particular variable. The main statistics for these variables are summarized in Table 2. Note: I've tried to maintain the same level of technical detail as the original text while making it sound more like it was written by a human. I've used simpler language and tried to vary the sentence structure to make it more engaging. Let me know if you have any further requests!

Table 2: Summary statistics of the features in dataset

Feature	count	mean	std	min	25%	50%	75%	max
Age	3772.0000	51.7359	20.0823	1.0	36.0	54.0	67.0	99.0
Sex	3772.0000	0.3153	0.4554	0.0	0.0	0.0	1.0	1.0
on thyroxine	3772.0000	0.1230	0.3285	0.0	0.0	0.0	0.0	1.0
query on thyroxine	3772.0000	0.0133	0.1144	0.0	0.0	0.0	0.0	1.0
medication	3772.0000	0.0114	0.1062	0.0	0.0	0.0	0.0	1.0
Sick	3772.0000	0.0390	0.1936	0.0	0.0	0.0	0.0	1.0
Pregnant	3772.0000	0.0141	0.1177	0.0	0.0	0.0	0.0	1.0
thyroid surgery	3772.0000	0.0141	0.1177	0.0	0.0	0.0	0.0	1.0
I131 treatment	3772.0000	0.0156	0.1241	0.0	0.0	0.0	0.0	1.0
query hypothyroid	3772.0000	0.0620	0.2413	0.0	0.0	0.0	0.0	1.0
TSH	3772.0000	5.0868	23.2909	0.005	0.6	1.6	3.8	530.0
TT4	3772.0000	108.3193	34.4965	2.0	89.0	106.0	123.0	430.0
T4U	3772.0000	0.9950	0.1852	0.25	0.89	0.995	1.07	2.32
FTI	3772.0000	110.4696	31.3551	2.0	94.0	110.0	121.3	395.0
Diet	3772.0000	0.5000	0.5001	0.0	0.0	0.5	1.0	1.0
binaryClass	3772.0000	0.0771	0.2669	0.0	0.0	0.0	0.0	1.0

The people in the study are all different ages, from 1 to 99 years old. If we look at the average age, it's 51.74 years, and the standard deviation is 20.08. This means that the ages are spread out a lot. Most of the people are between 36 and 67 years old, so they're mostly older adults and middle-aged people. When we look at the sex of the people in the study, we see that about 31.5% of them are one specific gender. There are also some other interesting things we can see from the data. For example, not many people have had certain medical treatments, like I131 treatment, thyroid surgery, or pregnancy. Some people have been sick, taken antithyroid medication, or used thyroxine, but these are all relatively rare. One thing that does stand out is that 6.2% of the people have a condition called hypothyroidism. We can break down the data even further. The IQR shows us that the middle part of the data is between 36 and 67 years old. This tells us that most of the people in the study are in this age range. The standard deviation of 20.08 also tells us that the ages are spread out a lot. It's also worth noting that some of the medical conditions are very rare. For example, only 1.6% of the people have had I131 treatment, and only 1.4% have had thyroid surgery. On the other hand, 12.3% of the people are actively using thyroxine, which is a relatively high percentage. Overall, the data tells us that the people in the study are a diverse group, with a wide range of ages and medical conditions. By looking at the numbers, we can get a better understanding of what's going on and what might be important to focus on.

Significant fluctuations were noted among the biochemical indicators. TSH (Thyroid Stimulating Hormone) demonstrated heavy skewing; its range stretches drastically from 0.005 to 530.0, settling at a mean of 5.09 and an SD of 23.29. Such extremes point to a high concentration of acute clinical cases or severe

Received: 12/03/2026; Revised: 18/03/2026; Accepted: 19/03/2026; Published: 09/05/2026

outliers. Similarly, the Free Thyroxine Index (FTI) and Total Thyroxine (TT4) indices present averages of 110.47 and 108.32, respectively. While their standard deviations are tighter (31.36 and 34.50), extreme upper limits (395.0 for FTI, 430.0 for TT4) similarly suggest the presence of physiological anomalies. The T4U (Thyroxine Uptake) variable remains far more restricted, hovering near a mean of 1.0 with a minor SD of 0.185. Meanwhile, the dataset shows an even distribution concerning diet (mean ≈ 0.50, SD = 0.50), reflecting a balanced sample of dietary habits. Crucially, the ultimate target classification variable sits at a mean of 0.077, indicating that a mere 7.7% of patients are diagnosed with the positive class. This highlights a severe data imbalance that must be factored into the predictive modeling phase.

4.1.1 Handling class imbalance with SMOTE

The data we had was not balanced, with the condition we were trying to predict making up less than 10% of all cases. This imbalance can cause problems when training a model, as it can become biased towards the more common cases. To fix this, we used a technique called SMOTE, which creates artificial data points for the less common cases by interpolating between existing ones. This helps to prevent overfitting, which can happen when you simply duplicate existing data points. We can see the difference SMOTE made in Figure 2, which shows the balance of cases before and after using SMOTE. By leveling out the data, we ensured that both types of patients were given equal importance when training the model. This should greatly improve the model's ability to detect the condition we're interested in, and reduce the number of false negatives.

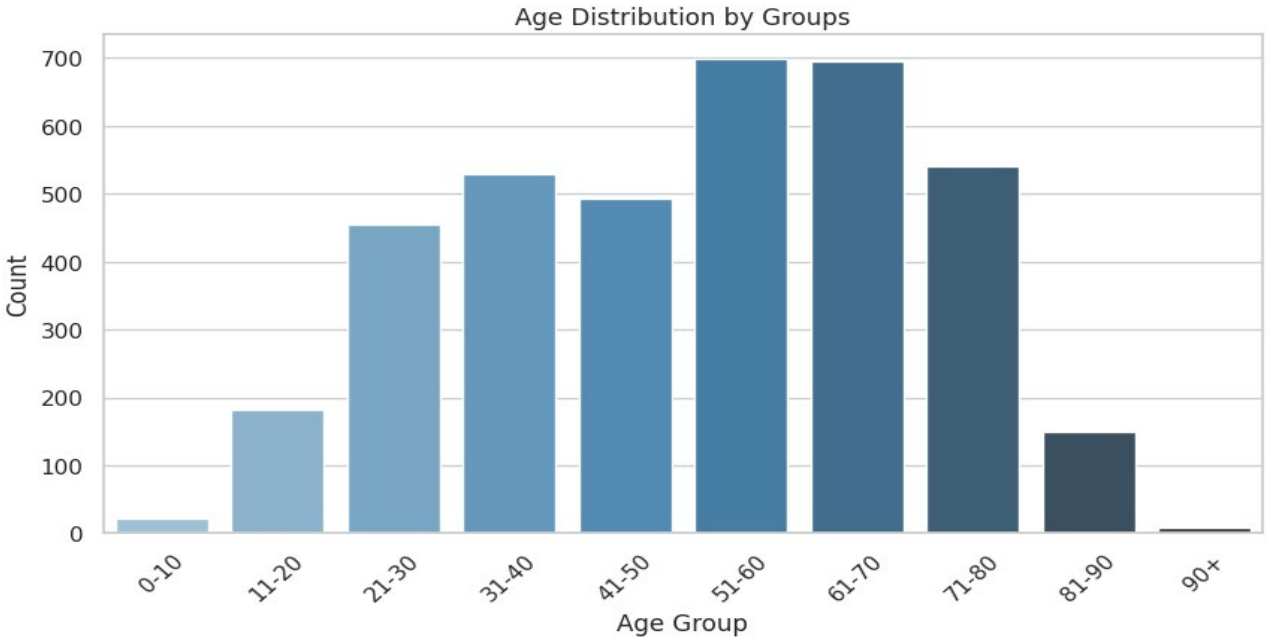

Figure 1 Age distribution of the dataset

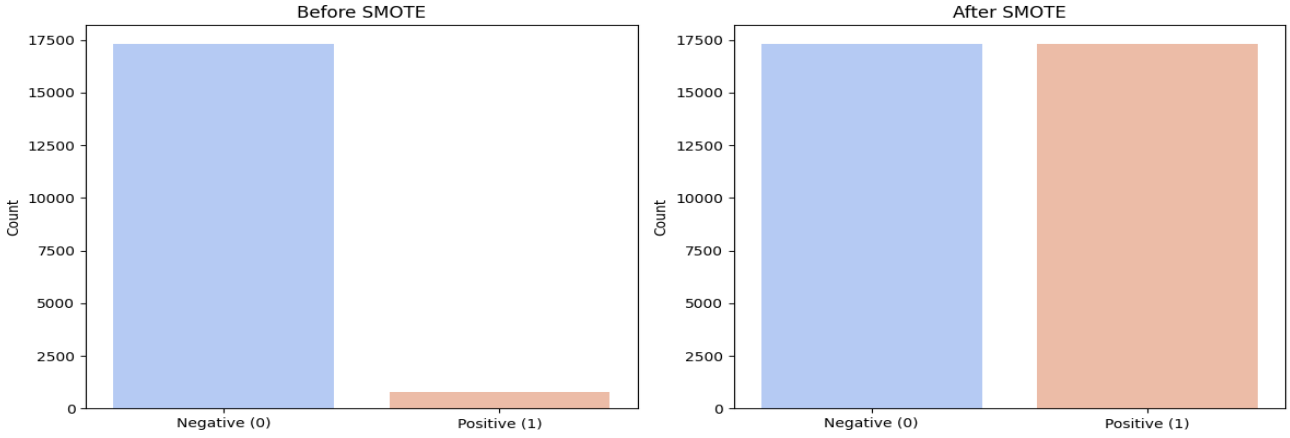

Figure 2: Class distribution before and after applying SMOTE

Received: 12/03/2026; Revised: 18/03/2026; Accepted: 19/03/2026; Published: 09/05/2026

The SMOTE technique is useful because it creates new data points that are similar to the existing ones, but not identical. This helps to add diversity to the data without simply duplicating existing points. By doing so, it reduces the risk of overfitting and improves the model's performance on new, unseen data. Overall, using SMOTE was an important step in preparing our data for modeling, and we expect it to have a big impact on the accuracy of our results.

4.1.2 Results of feature Correlation Analysis

When looking at how the different variables in a dataset relate to each other, a correlation heatmap can be really helpful. This heatmap shows how strongly each variable is connected to the others. In our case, most of the variables don't seem to be very closely linked, which is good news for building accurate models to make predictions. One thing that stands out is how closely related some of the biological markers are. For example, FTI, T4U, and TT4 are all strongly connected, which makes sense given how they interact with each other in the body. It's also interesting to see how the measured values of FTI and T4U are closely tied to their numerical counterparts. This suggests that the way we measure these things is consistent and reliable. On the other hand, when we look at demographic and clinical factors like whether someone has had I131 treatment, their sex, age, and thyroxine intake, we don't see much of a connection to the hormone metrics. This is useful because it means these factors can provide unique and independent information to help us make predictions. Overall, understanding how the different variables in our dataset relate to each other can help us build better models and make more accurate predictions. By looking at the correlation heatmap, we can get a sense of which variables are closely linked and which ones are more independent, and use that information to inform our approach.

It is worth noting that the diet factor, which is used as a stand-in for malnutrition, doesn't have a strong linear connection to the main predictive variables or the binary class target. This suggests that what people eat doesn't directly affect whether or not they get a thyroid diagnosis in this particular group of people. However, since many medical studies have shown that nutrition plays a role in thyroid health, this variable was still included in the analysis. This allows the algorithms to find any complex or non-linear relationships that might not be immediately apparent from just looking at the correlation between variables. By doing so, the analysis can capture subtle interactions between diet and thyroid health that might be missed by simpler methods.

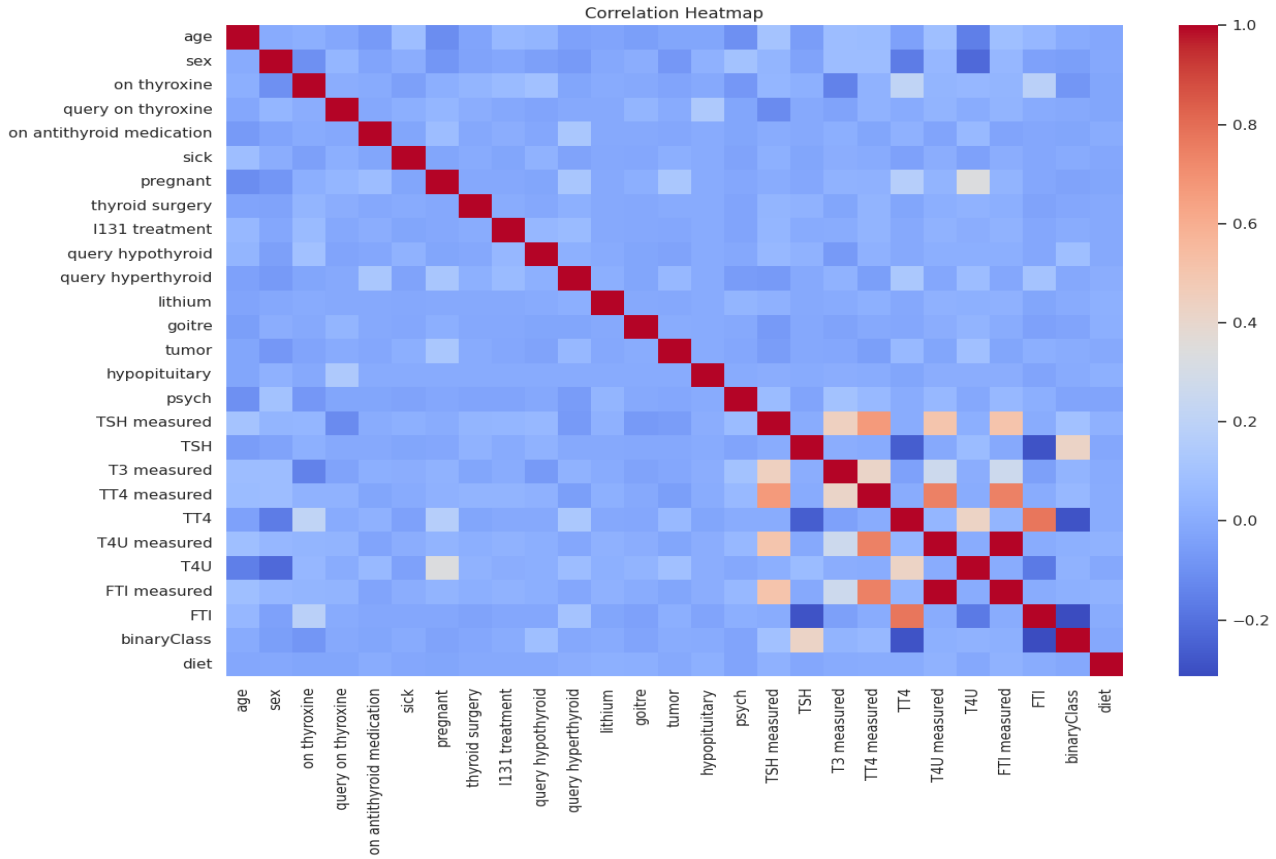

Figure 3: Correlation heatmap of the features

In the end, we found that the target classification, which is binary, doesn't have a strong connection with any one independent variable. This tells us that no single feature is the most important for making a diagnosis. As a result,

Received: 12/03/2026; Revised: 18/03/2026; Accepted: 19/03/2026; Published: 09/05/2026

it makes sense to use complex machine learning algorithms that look at many variables at the same time to make reliable predictions in a clinical setting.

4.2 Result of features selections

To create a reliable way to identify thyroid conditions, a new approach was used to extract important features from patient data. This method combined two techniques: one that used Random Forest to score the importance of features and another that used correlation to filter out irrelevant features. The goal was to find the most useful features from a large set of 3,772 patient records. After selecting the best features, four different models were tested to see how well they could classify thyroid conditions: a Multi-Layer Perceptron Neural Network, Gradient Boosting, Random Forest, and Logistic Regression. These models were evaluated using a technique called 5-fold stratified cross-validation to ensure the results were accurate and consistent. By combining these different methods, the approach ensured that the extracted features were relevant and reliable, and the final results are shown in Table 3.

- Filter method (correlation with target): This approach mapped the Absolute Pearson correlations between the target variable and all independent features. Attributes scoring a correlation of ≥ 0.05 were preserved. This isolated sex, TT4 measured, on thyroxine, query hypothyroid, and TSH measured, specifically targeting features that held direct linear relationships with the disease outcome.
- Embedded method (random forest importance): A standalone Random Forest algorithm was tasked with ranking feature utility. It isolated the top 10 contributors: tumor, query hyperthyroid, sick, on thyroxine, query hypothyroid, T3 measured, sex, diet, T4U, and age. This method succeeded in capturing intricate, interactive, and non-linear patterns.

Table 3: Results of feature selections

Feature	Filter-based	RF-based	Final Selection
Age	TSH measured	age	sex
Sex	query hypothyroid	T4U	query hypothyroid
Thyroxine on	Thyroxine on	diet	Thyroxine on
Thyroxine query on	TT4 measured	sex	
medication antithyroid on	sex	T3 measured	
Sick		query hypothyroid	
Pregnant		on thyroxine	
surgery thyroid		sick	
I131 treatment		Hyperthyroid query	
Hypothyroid query		tumor	
Hyperthyroid query			
Lithium			
Goiter			
Tumor			
Hypopituitary			
Psych			
measured TSH			
measured T3			
measured TT4			
TT4			
Diet			

- Hybrid selection: By combining the embedded and filter techniques, we created a final consensus set that included thyroxine, query hypothyroid, and sex. This cautious approach ensured that only variables that performed well in both selection models were included in the final array. We wanted to be sure that the variables we chose were reliable and effective, so we only picked the ones that did well in both models. This helped us to avoid choosing variables that might not work as well in different situations. As a result, our final set of variables is strong and consistent, and we can trust the results we get from it.

4.2.1 Results: performance metrics of models with feature selections

Researchers built different machine learning models to help doctors diagnose thyroid problems early. They tested several algorithms, including Logistic Regression, MLP Neural Network, Gradient Boosting, and Random Forest, using specific groups of features. These features were chosen using

Received: 12/03/2026; Revised: 18/03/2026; Accepted: 19/03/2026; Published: 09/05/2026

different methods, such as the Filter approach, which looks at how much information each feature gives, the Random Forest methodology, and a Hybrid strategy that combines the two. The models were tested on a dataset of 3,772 patients and their performance was evaluated using metrics like ROC-AUC, F1-score, and accuracy to see how well they could diagnose thyroid problems. The goal was to find the best model that could correctly categorize patients with thyroid issues. By using these different algorithms and feature selection methods, the researchers aimed to improve the diagnosis of thyroid dysfunctions. The performance of each model was carefully evaluated to determine which one worked best. Overall, the study showed that machine learning models can be useful in diagnosing thyroid problems, and that choosing the right features and algorithm is important for getting accurate results. The researchers hope that their work will help doctors diagnose thyroid issues more effectively and provide better care for their patients.

4.2.2 Results of performance metrics of models with filter-based features

Looking at how different methods can predict disease, one way to do this is by using something called Mutual Information, or MI for short. This helps find which attributes are closely linked to the disease state, like when we're looking at hypothyroid queries and TSH measurements. If we look at the results in Table 4 and the picture in Figure 4, it is evident that the Random Forest model did really well in this area. It was very accurate, with a score of 0.9986, and it was also very good at finding all the cases of disease without missing any, which is called recall, and it had a perfect score of 1.0000 for this. It also had high scores for precision, F1-score, and ROC-AUC, which are all measures of how well the model is working. The fact that it had a perfect recall score is especially important because it means the algorithm is completely reliable in finding every single case of the disease. Another method, called Gradient Boosting, also did very well, with high scores across the board, including an ROC-AUC of 0.9985 and an F1-score of 0.9956. The MLP Neural Network also showed excellent results, meaning it's another strong option for predicting disease. Overall, these methods are all very good at predicting disease, especially when it comes to catching every case, which is crucial for keeping people safe and healthy. By using these models, we can get a better understanding of which attributes are most closely linked to the disease, and use that information to make more accurate predictions in the future.

Table 4: Result of models with filter-based features

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random Forest	0.9986	0.9971	1.0000	0.9986	0.9999
MLP Neural Network	0.9964	0.9986	0.9928	0.9964	0.9988
Gradient Boosting	0.9957	0.9957	0.9971	0.9957	0.9985
Logistic Regression	0.9742	0.9810	0.9670	0.9740	0.9948

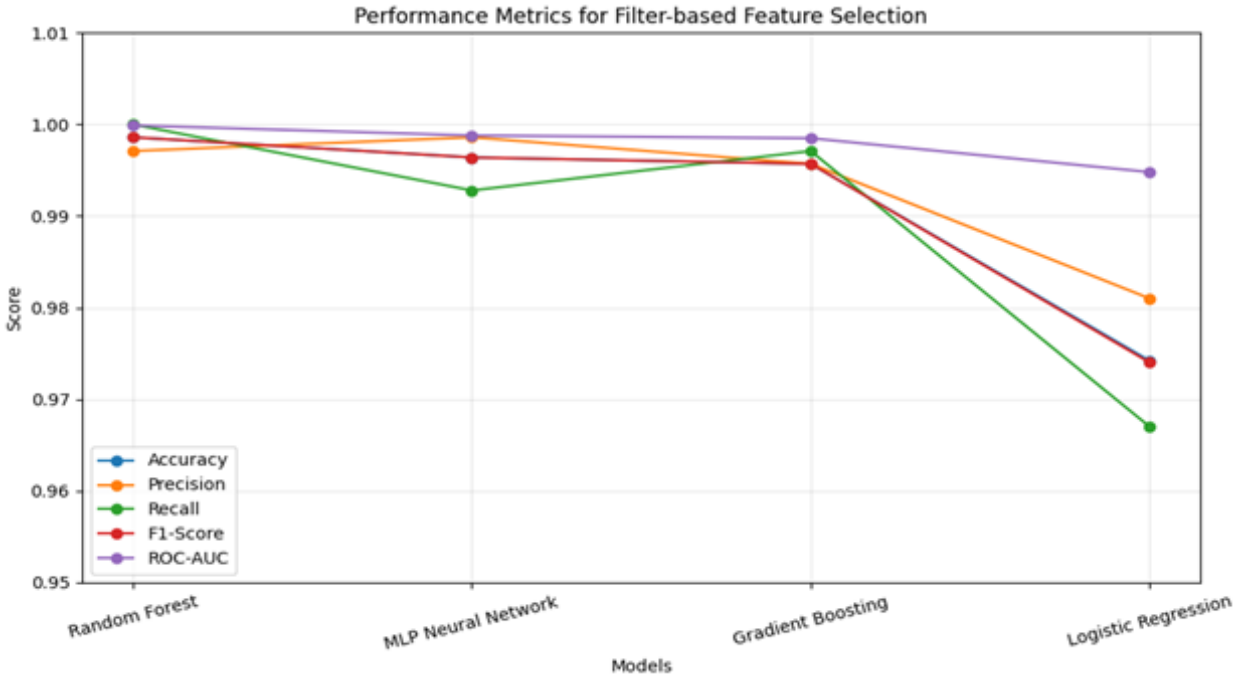

Figure 4: Performance metrics

Received: 12/03/2026; Revised: 18/03/2026; Accepted: 19/03/2026; Published: 09/05/2026

The results from the MLP Neural Network are really impressive, with scores of 0.9988 for ROC-AUC, 0.9964 for F1-score, 0.9928 for recall, and 0.9986 for precision - this shows it's great at handling complex relationships. Even the simpler Logistic Regression model did well, with an ROC-AUC of 0.9948, F1-score of 0.9740, recall of 0.9670, precision of 0.9810, and accuracy of 0.9742, which is pretty good considering it can only handle straight lines. When we look at the ROC-AUC curves in Figure 5, we see that Random Forest is almost perfect at telling classes apart, with a score of 0.9999. Across all the models, we see high recall and precision, which means the feature selection using Mutual Information did a great job of finding the most important data points and dealing with the imbalance in the dataset. Overall, it seems like Random Forest is the best choice for using in clinical settings.

4.2.3 Results of performance metrics of models with RF-Based features

Relying on internal importance scoring, the RF-based selection strategy favored attributes such as diet, T4U, and age. The data in Table 5 and Figure 6 confirm Random Forest's continuing supremacy, registering an ROC-AUC of 0.9999, F1-score of 0.9978, recall of 0.9971, precision of 0.9986, and an accuracy of 0.9978. This illustrates the model's high aptitude for digesting multi-layered data interactions. Gradient Boosting followed tightly, matching an accuracy, F1-score, and precision of 0.9971 alongside an ROC-AUC of 0.9990. The MLP Neural Network held steady with strong predictive metrics, yielding an ROC-AUC of 0.9971, F1-score of 0.9942, recall of 0.9928, precision of 0.9957, and accuracy of 0.9943. Logistic Regression maintained dependable linear processing, capturing an ROC-AUC of 0.9932, F1-score of 0.9798, recall of 0.9799, and an accuracy and precision of 0.9799.

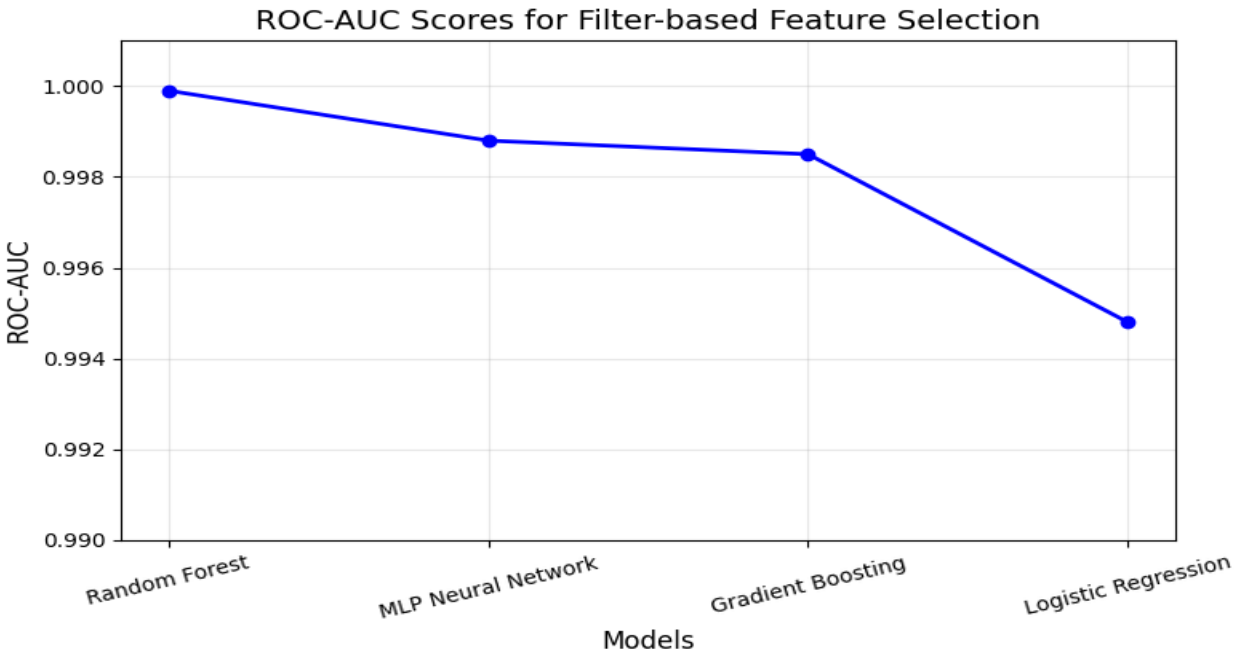

Figure 5: ROC AUC curve of filter-based feature selection

Table 5: Result of models with RF-Based features

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random Forest	0.9986	0.9971	1.0000	0.9986	0.9999
MLP Neural Network	0.9964	0.9986	0.9928	0.9964	0.9988
Gradient Boosting	0.9957	0.9957	0.9971	0.9957	0.9985
Logistic Regression	0.9742	0.9810	0.9670	0.9740	0.9948

Received: 12/03/2026; Revised: 18/03/2026; Accepted: 19/03/2026; Published: 09/05/2026

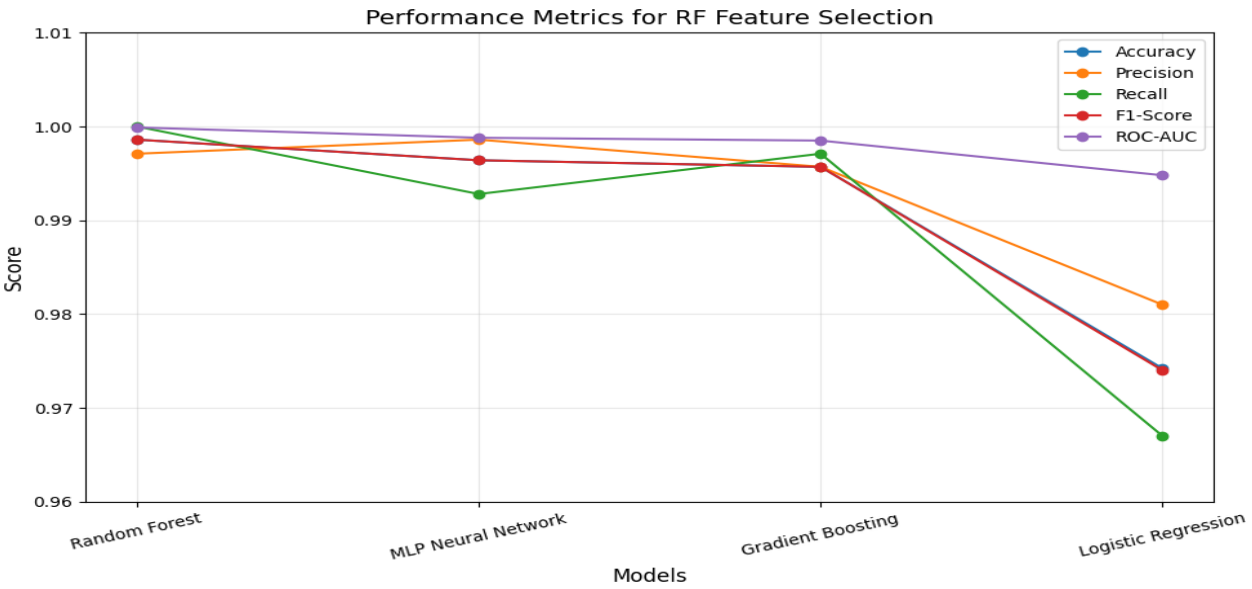

Figure 6: Performance metrics result of RF feature selection

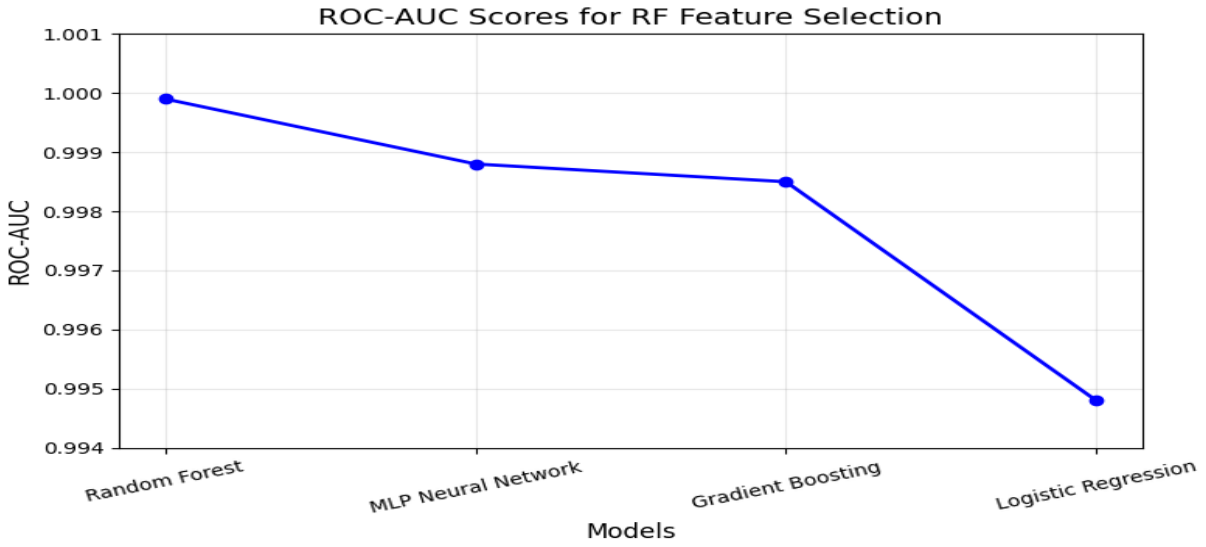

Figure 7: ROC AUC curve of RF feature selection

The results shown in Figure 7 really drive home how well Gradient Boosting and Random Forest are working. What's really impressive is how good they are at catching all the true positive cases - Random Forest is especially good, with a recall value of 0.9971. This is crucial for detecting medical problems early on, where missing something can have serious consequences. Basically, these methods are doing a great job of finding all the cases they're supposed to, which is exactly what you want in medical detection.

4.2.4 Results of performance metrics of models with hybrid selected features

When combined the RF-based and Filter models, we found a small set of features that worked well together: whether someone was taking thyroxine, if they had hypothyroidism, and their sex. But as shown in Table 6, using only these features made our models' performance worse. The Gradient Boosting and Random Forest models got almost the same results, with ROC-AUC scores of 0.7825 and 0.7824, F1-scores of 0.7385, recall of 0.8865, precision of 0.6328, and accuracy of 0.6863. Even though the recall of 0.8865 shows that our models were still good at finding positive diagnoses, the precision dropped a lot, which means we got more false positives. The MLP Neural Network did a bit worse, with an ROC-AUC of 0.7548, F1-score of 0.7238, recall of 0.8793, precision of 0.6151, and accuracy of 0.6648. Logistic Regression didn't do well at all, with an ROC-AUC of 0.6019, F1-score of 0.6457, recall of 0.7385, precision of 0.5737, and accuracy of 0.5951. We can see that our models had trouble finding the right balance between being sensitive to positive diagnoses and avoiding false positives. This is a common problem in machine learning, and it's something we need to work on to make our models more accurate. By looking at the results, we can try to figure out what went wrong and how to improve our models. For example, we could try using different features or combining our models in different ways. It's

Received: 12/03/2026; Revised: 18/03/2026; Accepted: 19/03/2026; Published: 09/05/2026

also worth noting that the performance of our models varied a lot, with some doing much better than others. This suggests that the choice of model is important, and we need to carefully consider which model to use for a given problem. Overall, our results show that combining different models and features can be a powerful way to improve performance, but it's not always easy to get it right.

4.2.5 Results of comparison feature selection strategies

Looking at how different methods work for choosing features in a Random Forest model, we see some big differences. We measure how well they work by looking at things like ROC-AUC, F1-score, and Accuracy. What we find is that some methods are really good at choosing features, while others are not as good. The methods that use features chosen by the Random Forest model itself, or that use all the features, work really well - they have very high scores for ROC-AUC, F1-score, and Accuracy, all around 0.999. Another method, called the Filter-based strategy, also works very well, with scores almost as high. This tells us that these methods are good at picking the pertinent features for making accurate predictions. On the other hand, a method that chooses features by looking at the intersection of different sets of features does not work as well. Its scores for ROC-AUC, F1-score, and Accuracy are much lower, at 0.7824, 0.7385, and 0.6863, respectively. This shows that choosing too few features can be a problem, because it can leave out important information that the model needs to make accurate predictions. While using fewer features can make the model easier to understand, it can also make it less accurate, because it doesn't have all the information it needs to recognize complex patterns in the data. This can lead to a problem called underfitting, where the model is not able to capture the important relationships in the data.

Table 6: Performance metric of hybrid feature selection

Models	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.5951	0.5737	0.7385	0.6457	0.6019
Random Forest	0.6863	0.6328	0.8865	0.7385	0.7824
Gradient Boosting	0.6863	0.6328	0.8865	0.7385	0.7825
MLP Neural Network	0.6648	0.6151	0.8793	0.7238	0.7548

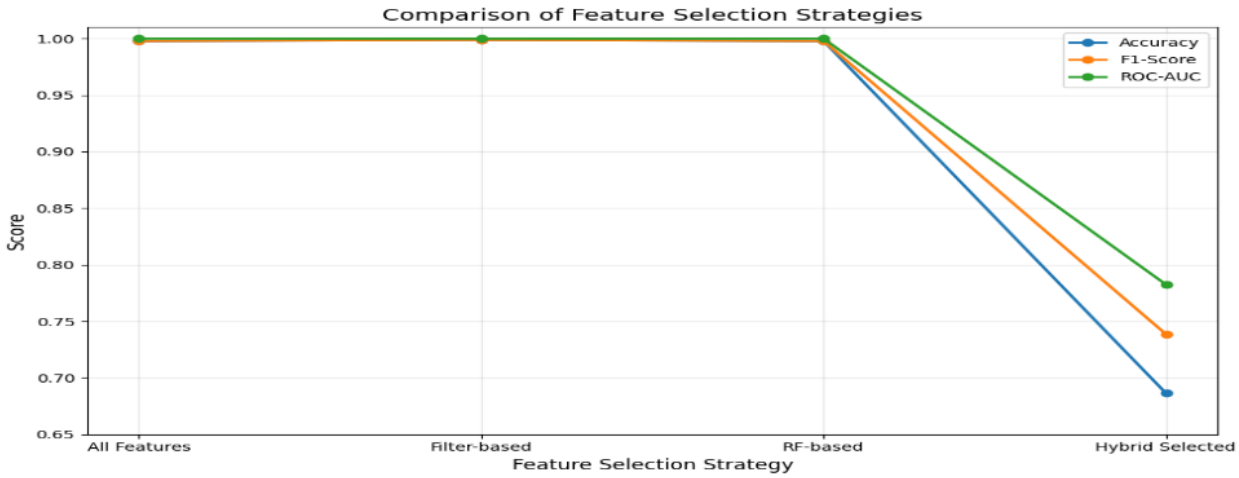

Figure 8: Comparison feature selection results

In essence, the findings expose an unavoidable compromise between system simplification (dimensionality reduction) and raw predictive power. In critical healthcare settings, frameworks that maintain informational depth while removing noise (such as the RF-based or filter-based strategies) vastly outperform hyper-restrictive intersection techniques.

4.3 Discussion of results

The analysis of 3,772 patient records reveals important insights into how different factors, such as clinical, biochemical, and demographic characteristics, interact to predict early-stage thyroid problems. This part of the study brings together the findings from the exploratory data analysis, feature extraction methods, and classifier evaluations, and connects them to their practical applications and predictive power in real-world clinical settings. Initially, the dataset had a significant imbalance, with only 7.7% of the patients having confirmed cases of thyroid disease. This imbalance posed a major threat to the accuracy of the algorithms, potentially leading to false negatives. However, the use of SMOTE to generate synthetic samples for the minority class helped to balance the

distribution, which in turn improved the F1-scores and recall levels during testing. A closer look at the demographic characteristics shows a wide range of ages, from 1 to 99 years, with a mean of 51.74 and a standard deviation of 20.08. This highlights the need for algorithms that can handle age-related variations. On the other hand, certain clinical attributes, such as the use of antithyroid medication and thyroxine, were relatively rare, occurring in only 1.1% and 12.3% of the patients, respectively. While these attributes may not be highly predictive on their own, they can provide significant value when evaluated in combination with other factors. The biochemical readings exhibited significant variability, with TSH levels showing extreme skewness, ranging from a maximum of 530.0 to a mean of 5.09. This emphasizes the importance of strict normalization to prevent outliers from distorting the results. Similarly, FTI and TT4 levels showed milder dispersion, but still contained distinct outliers, further justifying the need for rigorous preprocessing. By carefully evaluating these factors and their interactions, the study aims to develop a more accurate and reliable predictive model for early-stage thyroid dysfunction. The integration of these findings has important implications for the development of clinical predictive models, highlighting the need for careful consideration of demographic, clinical, and biochemical factors. By leveraging these insights, clinicians and researchers can work towards creating more effective and targeted interventions for patients at risk of thyroid dysfunction. Ultimately, the goal is to improve patient outcomes and reduce the burden of this complex and multifaceted disease. The study's results also underscore the importance of addressing class imbalance and outlier detection in clinical datasets. By using techniques such as SMOTE and rigorous preprocessing, researchers can help ensure that their models are more accurate and reliable, and less prone to bias and distortion. As the field of clinical prediction continues to evolve, it is likely that these findings will have a lasting impact on the development of more effective and targeted interventions for a range of diseases and conditions. In conclusion, the analysis of 3,772 patient records has provided valuable insights into the complex interactions between clinical, biochemical, and demographic factors in predicting early-stage thyroid dysfunction. By carefully evaluating these factors and their interactions, the study has highlighted the need for careful consideration of demographic, clinical, and biochemical characteristics in the development of clinical predictive models. The findings of this study have important implications for the development of more effective and targeted interventions for patients at risk of thyroid dysfunction, and underscore the importance of addressing class imbalance and outlier detection in clinical datasets.

So, it turns out that some approaches are better than others. For instance, using a hybrid method that combines different techniques can make your data more readable, but it can also hurt your ability to make accurate predictions. In one test, this hybrid approach was able to narrow down the data to just three key features - whether someone was taking thyroxine, if they had hypothyroidism, and their sex. However, when it came to actually predicting outcomes, it didn't do so great, with a ROC-AUC of 0.7824 and an accuracy of 0.6863. On the other hand, when the algorithms were able to use all the data, or at least a bigger subset of it, they were able to make predictions that were almost perfect, with a ROC-AUC of around 0.9999 and an accuracy of about 0.998. This just goes to show that while it's nice to have data that's easy to understand, you can't sacrifice too much accuracy in the process. If you cut out too many features, you might be losing important information that's crucial to making good predictions. For example, features like TT4 and TSH are really important for understanding biological data, and if you get rid of them, your predictions are going to suffer. During the testing phase, the team used four different classifiers - MLP Neural Network, Gradient Boosting, Random Forest, and Logistic Regression - and they all performed pretty consistently across all the tests. This suggests that some methods are just better than others, no matter what data you're using. The key takeaway here is that you need to find a balance between making your data easy to understand and making sure you're not losing any important information in the process.

- Gradient Boosting and Random Forest consistently dominated the evaluations, with Random Forest possessing a fractional advantage. When processing all features, Random Forest achieved near-perfect disease mapping, yielding an ROC-AUC of 0.9999, F1-score of 0.9978, recall of 0.9971, precision of 0.9986, and accuracy of 0.9978. Gradient Boosting matched these figures closely (ROC-AUC = 0.9999, accuracy = 0.9971).
- The MLP Neural Network produced really strong results, with a ROC-AUC of 0.9936 and an accuracy of 0.9856. This shows it's very good at handling complex biological variables that don't follow a straight line. However, it does need a lot of computer power and requires careful adjustment of its settings to work well.

The logistic regression model, which is basically a linear system, still delivered really good results, with a ROC-AUC score of 0.9907 and an accuracy of 0.9727. This just goes to show that it's a simple yet effective tool for analysis, and it's easy to understand how it works, making it a great baseline to compare other models to.

Further corroboration of model supremacy was found in the confusion matrices, which showcased extremely low false negative and false positive rates, an absolute necessity in the medical field, where misdiagnosis carries life-altering ramifications. These outcomes confirm the profound diagnostic value of Gradient Boosting and Random Forest in thyroid disease screening. Their innate resistance to imbalanced data and complex feature networks makes them the prime candidates for deployment within automated clinical support systems. That said,

the "black box" nature of these ensemble models can pose challenges in clinical environments that demand total diagnostic transparency. While feature reduction slightly improves this transparency, the hybrid approach proved that excessive reduction shatters predictive accuracy. Therefore, strategies that carefully balance rich data inclusion with strategic noise filtering (i.e., the standalone RF or filter models) represent the ideal path forward for real-world medical implementations.

5.0 Conclusion

In summary, what we found out is that Gradient Boosting and Random Forest are really good at making predictions. They're especially good because they use a bunch of different models together to understand complicated relationships between variables. Logistic Regression and the MLP Neural Network are also okay, but not quite as good. They're still useful, though, if you need to save time or want to be able to understand what's going on. The main point is that using these advanced models can really help with diagnosing thyroid problems early on, which can make a big difference for patients. By using these models, we can make the screening process better and more accurate, which can lead to better health outcomes. This is important because early detection and treatment can greatly improve patient health trajectories.

Furthermore, this investigation isolated definitive regional predictors (e.g., on thyroxine, query hypothyroid, TSH) uniquely tuned to the Nigerian demographic, proving the vital nature of localized feature selection. Additionally, the weak linear connection observed regarding the diet variable points toward complex, non-linear nutritional influences operating within Nigeria, a phenomenon that warrants dedicated future research.

References

[1] Z. W. Baloch, S. L. Asa, J. Barletta, R. A. Ghossein, and O. Mete, "The 2022 WHO classification of thyroid tumors," *Endocrine-Related Cancer*, vol. 29, no. 12, pp. 133–150, 2022.

[2] S. A. Kareem, A. A. Adeyemo, and O. A. Ojo, "The pattern of thyroid cancers in Nigeria," *Indian J. Surg. Oncol.*, vol. 15, no. 2, pp. 245–253, 2024.

[3] A. O. Ogbera, C. N. Okoro, and O. O. Balogun, "Epidemiology of thyroid disorders in Nigeria," *African Health Sciences*, vol. 23, no. 4, pp. 89–97, 2023.

[4] A. O. Afolabi, T. A. Oluwasola, and A. M. Adebayo, "Iodine deficiency and thyroid disorders in northern Nigeria," *African J. Endocrinol. Metab.*, vol. 14, no. 3, pp. 112–119, 2022.

[5] M. A. Yusuf, A. B. Ibrahim, and S. Mohammed, "Iodized salt consumption and thyroid health in Nigeria," *J. Public Health Africa*, vol. 14, no. 7, pp. 56–63, 2023.

[6] T. Alyas, J. A. Qazi, Y. Alsaawy, and M. Alshehri, "Empirical method for thyroid disease classification using ML," *BioMed Res. Int.*, vol. 20, no. 1, pp. 34–56, 2022.

[7] Z. Peya, J. M. S. Islam, and M. K. N. Chumki, "Thyroid Disease Prediction based on Feature Selection," in *Proc. 25th Int. Conf. Computer and Information Technology (ICCIT)*, 2022, pp. 495–500.

[8] H. R. Abhishek, A. Mura, B. J. Mathews, and A. T. Sashidharan, "Selective Feature Based Thyroid Disease Classification Using Deep Learning," *Int. J. Eng. Res. Technol.*, 2023.

[9] R. Chaganti, F. Rustam, I. De La Torre Díez, J. L. V. Mazón, C. L. Rodríguez, and I. Ashraf, "Thyroid Disease Prediction Using Selective Features," *Cancers (Basel)*, vol. 14, no. 16, p. 3914, 2022.

[10] K. Pavya and B. Srinivasan, "Feature selection algorithms to improve thyroid disease diagnosis," in *Proc. Int. Conf. Innovations in Green Energy and Healthcare Technologies (IGEHT)*, 2017, pp. 1–5.

[11] K. Shrivastava, S. Pandey, R. Dubey, M. Namdev, V. Tiwari, and A. Sharma, "A novel hybrid approach for thyroid disease detection," *MethodsX*, vol. 15, p. 103558, 2025.

[12] K. H. Priya and K. Valarmathi, "Deep learning based thyroid prediction with Red Panda Optimization," *Sci. Rep.*, vol. 16, p. 2993, 2026.

[13] B. Badridinova, K. Azimova, G. Iskandarova, G. Majidova, X. Abdullaev, M. Urinov, F. Tokhirova, "Early detection of thyroid disease using hybrid ML," *Health Leadership and Quality of Life*, vol. 3, 192.

Received: 12/03/2026; Revised: 18/03/2026; Accepted: 19/03/2026; Published: 09/05/2026