iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VII **Month of publication:** July 2026

DOI: https://doi.org/10.22214/ijraset.2026.84008

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

Data-Driven Exploration of Thyroid Disorders: An Exploratory Analysis of Clinical and Demographic Features

Parashuram S. Vadar[1], Urmila R. Pol[2]
YCSRD, Department of Computer Science, Shivaji University, Kolhapur, Shivaji University, Kolhapur,
[1]parashuramvadar@gmail.com

Abstract: The healthcare sector generates an enormous volume of data, making its management and analysis a significant challenge. Machine learning techniques have emerged as effective tools for processing and interpreting such large-scale medical datasets. According to the National Family Health Survey (NFHS-5), the prevalence of thyroid disorders is steadily increasing in India, with approximately one in every ten adults affected. It is estimated that over 42 million people in the country are living with thyroid-related diseases. Accurate analysis of clinical data is essential for the timely and precise diagnosis of these conditions. In this study, thyroid disorders were categorized into five classes: Euthyroid, Overt Hypothyroidism, Overt Hyperthyroidism, Subclinical Hyperthyroidism, and Subclinical Hypothyroidism. The primary objective of this research is to perform an exploratory data analysis (EDA) of the thyroid dataset to uncover meaningful patterns and relationships among hormone levels, age, and gender. The findings from the EDA highlight significant gender-based differences, age-related variations, and important correlations among thyroid hormone parameters, providing valuable insights for thyroid disease assessment.
Keywords: Thyroid, Machine Learning, Euthyroid, Hyperthyroidism, Hypothyroidism, Exploratory Data Analysis.

I. INTRODUCTION

Thyroid disorders denote an important subset of endocrine diseases which are most frequent in India. These disorders are the second most common in endocrine pathology. Worldwide approximately 8-20% of adult population are affected with thyroid-related diseases, out of which, less than 5% are diagnosed with cancer. The load of thyroid disorders in the overall population is massive. The primary reason for the onset of these diseases is the malfunction of the thyroid gland. A recent report estimates that 300 million people suffer from thyroid disorders worldwide, including 42 million Indians. This chronic non-communicable disease affects more women than men. The probability of women getting thyroid disease is five to eight times higher than that of men.

II. REVIEW OF LITERATURE

Research papers focusing on the Exploratory Data Analysis (EDA) of thyroid disease primarily analyze data patterns, clinical features, and demographic correlations before feeding them into diagnostic machine learning systems. Most studies use the iconic UCI Machine Learning Repository Thyroid Dataset or localized clinical records to map biomarker behaviours. The following peer-reviewed and indexed research papers highlight the role of data exploration, feature analysis, and data mining in diagnosing thyroid disease:

This paper investigates structural data profile and class distributions (finding a breakdown of 8.05% Hypothyroid vs 2.49% Hyperthyroid vs healthy controls). Comprehensive exploratory screening confirms that Thyroid-Stimulating Hormone (TSH) serves as the single most critical indicative biomarker during early visualization [1].

Exploratory Data Analysis in this study concentrates on data visualization, categorical feature encoding, missing value profiling, and diagnosing clinical target imbalances. Maps out correlations between physical attributes and laboratory metrics, showing that optimizing the EDA preprocessing phase drastically minimizes errors in downstream ensemble models [2].

This study evaluates the distribution shifts of physiological features and targets feature redundancy reduction using the UCI Repository. It uses EDA alongside SHAP (SHapley Additive exPlanations) to prove that narrowing down to just ten consensus clinical features (led by TSH and FTI) preserves a near-perfect diagnostic accuracy of 99.7% [3].

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

It explores correlations between patient demographics (age, gender), clinical records (thyroid surgery, medication status), and direct blood hormone measurements. Demonstrates how data cleaning and exploratory correlation maps help fine-tune advanced baseline models like CatBoost and Random Forest [4].
It applies rigorous exploratory data analysis techniques to address small sample sizes and highly skewed classification distributions. The exploration sets up a Bayesian Network framework that successfully integrates expert clinical domain knowledge with feature interactions to prevent overfitting [5].
It focuses on a localized dataset (the HAPPY study), using exploratory network analysis and community detection to uncover early-pregnancy symptom distributions. Maps unique symptom networks for expectant mothers, revealing that clusters containing fluid retention, memory loss, and chronic fatigue are heavily correlated with clinical thyroid impairment [6].
This study conducts a broad exploratory screening of UCI dataset variables, testing outlier frequencies across multiple algorithms (like LightGBM, Extra-Trees, and KNN. Pinpoints abnormal threshold boundaries in exploratory histograms, enabling early therapeutic prescription pipelines [7].
It explores early-stage clinical patterns using statistical mining to capture latent correlations before severe clinical symptoms show up. Proves that hybrid feature extraction routines rooted in preliminary data distribution scans maximize ensemble intelligence accuracy [8].
It evaluates demographic data profiling across thousands of patients, comparing distinct distributions between male and female clinical attributes. Visually outlines how female cohorts bear a higher statistical weight in dataset samples, which directly influences treatment planning and target thresholds [9].
Focuses on the automated profile evaluation of the clinical data elements in the Garavan Institute dataset using low-code pipelines. Establishes that early variance and covariance modeling during automated exploration directly improves the setup and deployment of classification models [10].

III. METHODOLOGY

Machine learning is used to quickly and accurately diagnose thyroid diseases because it has become an important tool in the health field and helps us diagnose and group diseases. For this reason, researchers gathered a lot of first-hand information from hospitals about thyroid patients who had different types of thyroid diseases like EUTHYROID, Overt Hypothyroidism, Overt Hyperthyroidism, Subclinical Hyperthyroidism, and Subclinical Hypothyroidism respectively. The thyroid dataset contains 1294 samples with 8 attributes namely Gender, Age, Location U/R, T3, T4, TSH, Free T4, Target. Various data pre-processing techniques were applied on dataset.

IV. EXPLORATORAY DATA ANALYSIS

EDA is an iterative process that involves exploring different aspects of the data, refining analyses, and gaining deeper insights. Prior to conducting additional analysis or constructing predictive models, it aids in comprehending the organization, trends, and connections within the data. It helps in understanding the data's characteristics, detecting anomalies or biases, and forming a solid foundation for subsequent data analysis and modeling tasks. EDA plays a crucial role in making informed decisions, uncovering patterns, and extracting meaningful information from the dataset.

A. Univariate Analysis

Each feature is evaluated separately. Pandas describe feature to produce summary statistics for all variables in tabular format. Such statistics are also useful to clarify issues.

B. Distribution Of Thyroid Disease Subtype

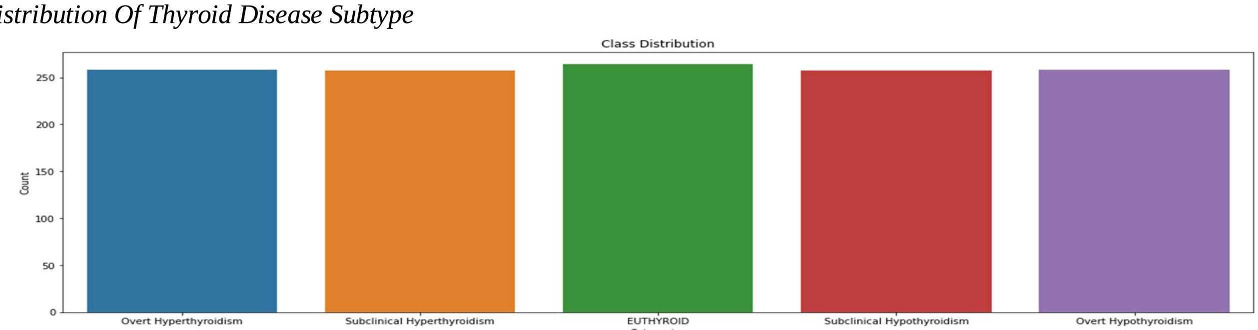


Fig. 1. Balanced Classes in the dataset

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

The visual representation of Target class as a balanced occurrence among the various types of thyroid diseases considered. Notably, 'EUTHYROID' exhibits a slightly higher prevalence compared to other subtypes, while 'Overt Hypothyroidism,' 'Overt Hyperthyroidism,' 'Subclinical Hyperthyroidism,' and 'Subclinical Hypothyroidism' display similar frequencies. This equitable distribution among subtypes could profoundly influence the research focus, encouraging a deeper investigation into distinct characteristics or attributes associated with each subtype.

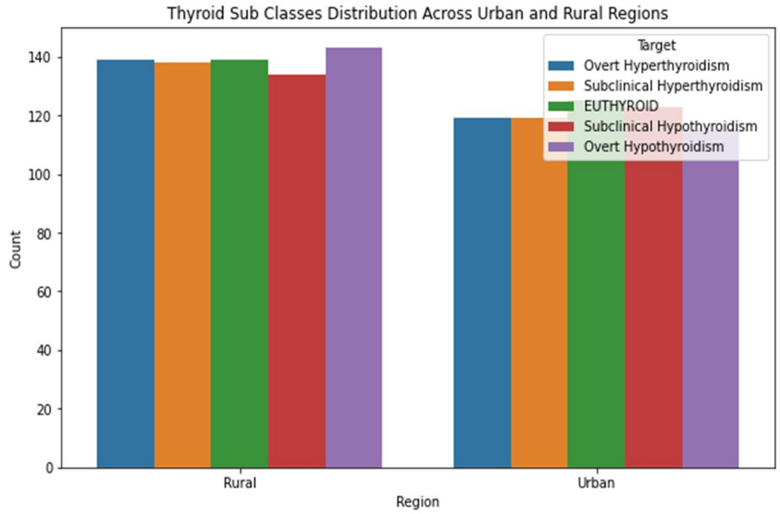


Fig. 2. Thyroid Sub Classes Distribution Across Urban and Rural Regions

The count plots for thyroid subclasses across rural and urban regions illustrate in the above figure. It is observed that, thyroid subclasses in rural areas shows higher counts for specific subclasses (potentially indicated by the taller bars in certain categories), suggesting a relatively higher incidence or prevalence of those subclasses in rural regions. Conversely, the count plot for urban areas displays relatively lower counts for those same subclasses compared to rural areas.

C. Distribution Statistics Of Age Feature

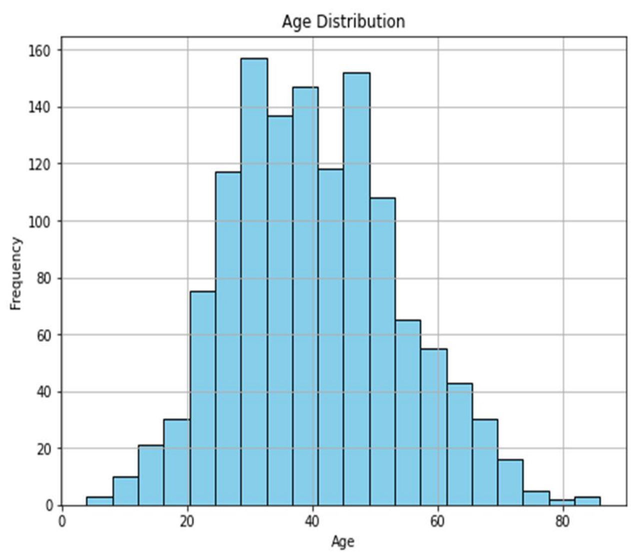


Fig. 3. Distribution Plot for Age feature

The analysis of the age distribution within the dataset of individuals diagnosed with thyroid disease reveals a notable concentration of cases between the ages of 30 to 50 years old. This finding underscores a significant trend wherein a majority of individuals diagnosed with thyroid-related issues fall within this specific age bracket.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

D. Distribution For Continuous Data In The Dataset

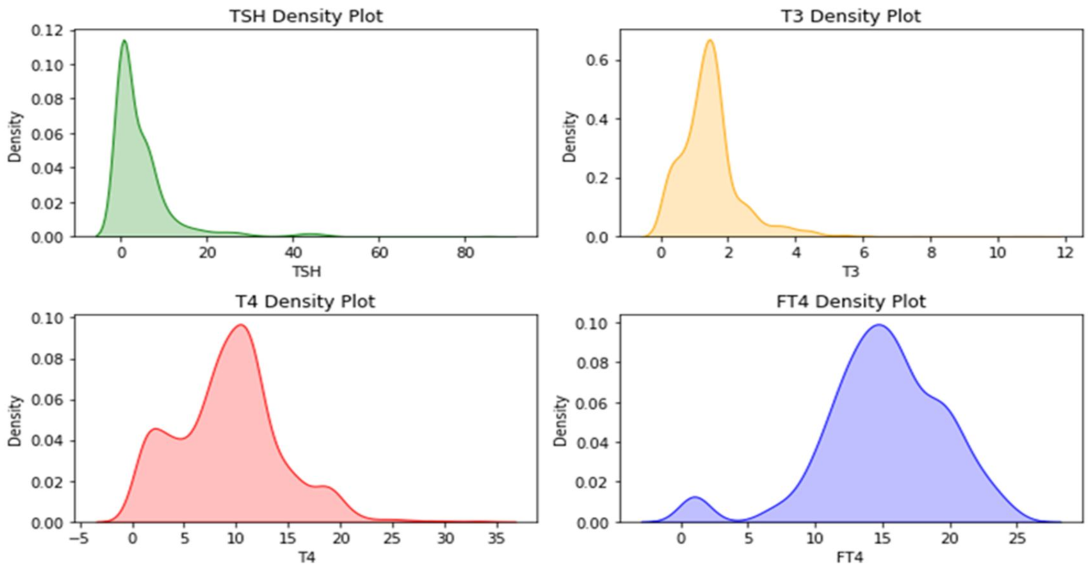


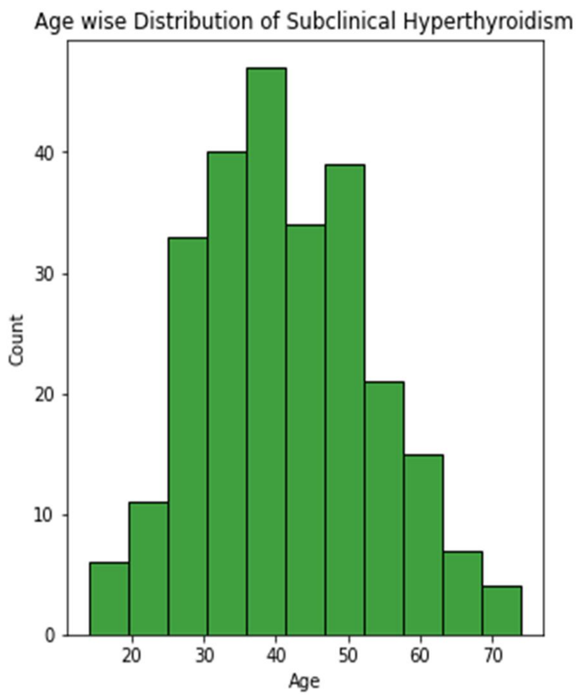
Fig. 4. Density Plots of TSH, T3, T4 and FT4

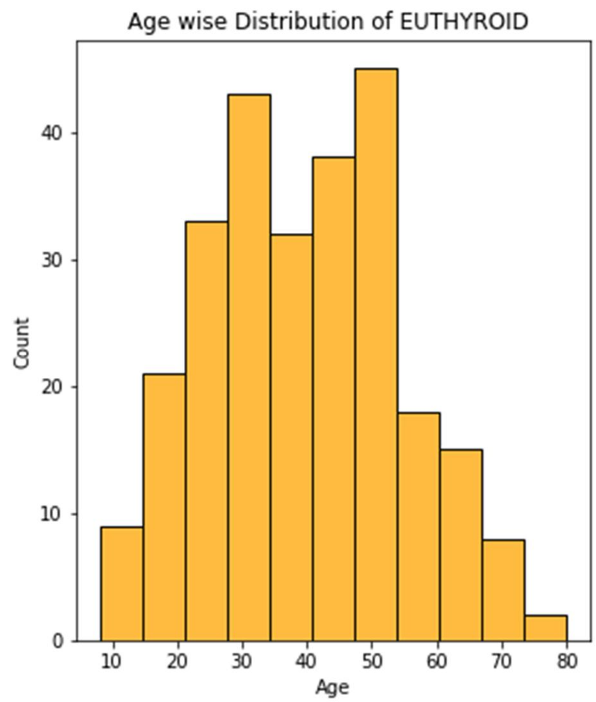
It is observed from the density plot that the values in the dataset for TSH, T3, T4 and FT4 looks heavily skewed towards left.

E. Multivariate Analysis

Multivariate analysis (MVA) is a statistical approach that involves the observation and analysis of multiple variables and their outcomes simultaneously, based on principles of multivariate statistics. Multivariate Variate Analysis (MVA) is commonly employed to address scenarios where several measurements are conducted on each experimental device, and the associations between these measurements and their structures are significant.

F. Age Wise Distribution Of Thyroid Sub Classes





ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

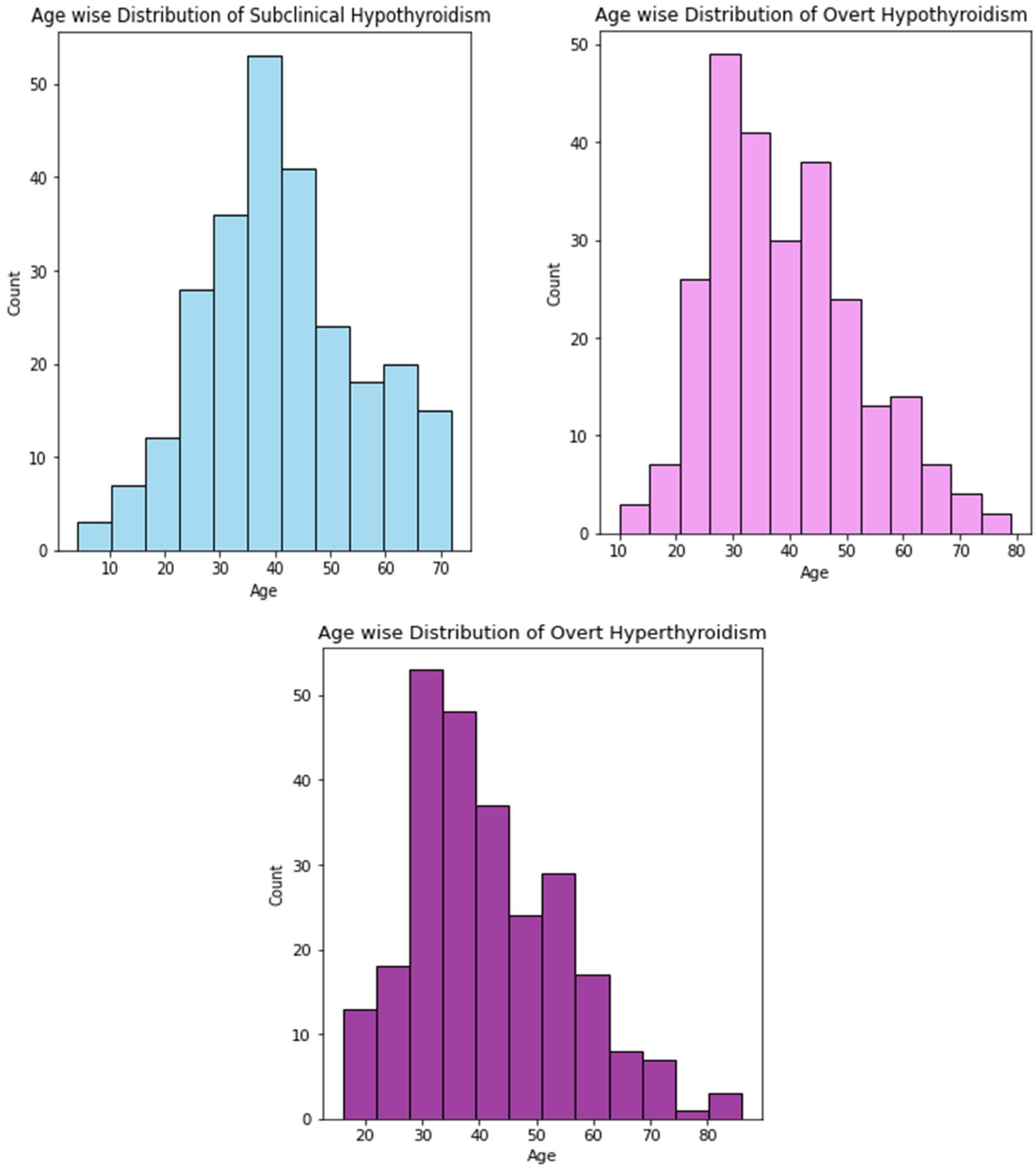


Fig. 5. Frequency Distribution of Thyroid Sub Types with Age

The analysis of thyroid sub conditions across different age groups reveals distinct patterns in the prevalence of thyroid disorders. Overt Hyperthyroidism predominantly affects individuals between the ages of 30 and 40, indicating a concentrated occurrence within this relatively narrow age range. Subclinical hyperthyroidism exhibits a broader age distribution, with a peak between 30 and 40 but extending up to the age of 55, suggesting a sustained presence across a wider age group. Normal thyroid function, referred to as Euthyroid, is prevalent across a broad age range, spanning from 25 to 55. This indicates that the majority of individuals in the dataset exhibit normal thyroid function throughout adulthood. Subclinical Hypothyroidism is more concentrated within a specific age range, primarily affecting individuals between 35 and 45. Similar to subclinical hypothyroidism, overt hypothyroidism is concentrated within a relatively narrow age range from 25 to 45, indicating a significant prevalence within this specific age group.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

G. Gender Wise Distribution Of Thyroid Sub Classes

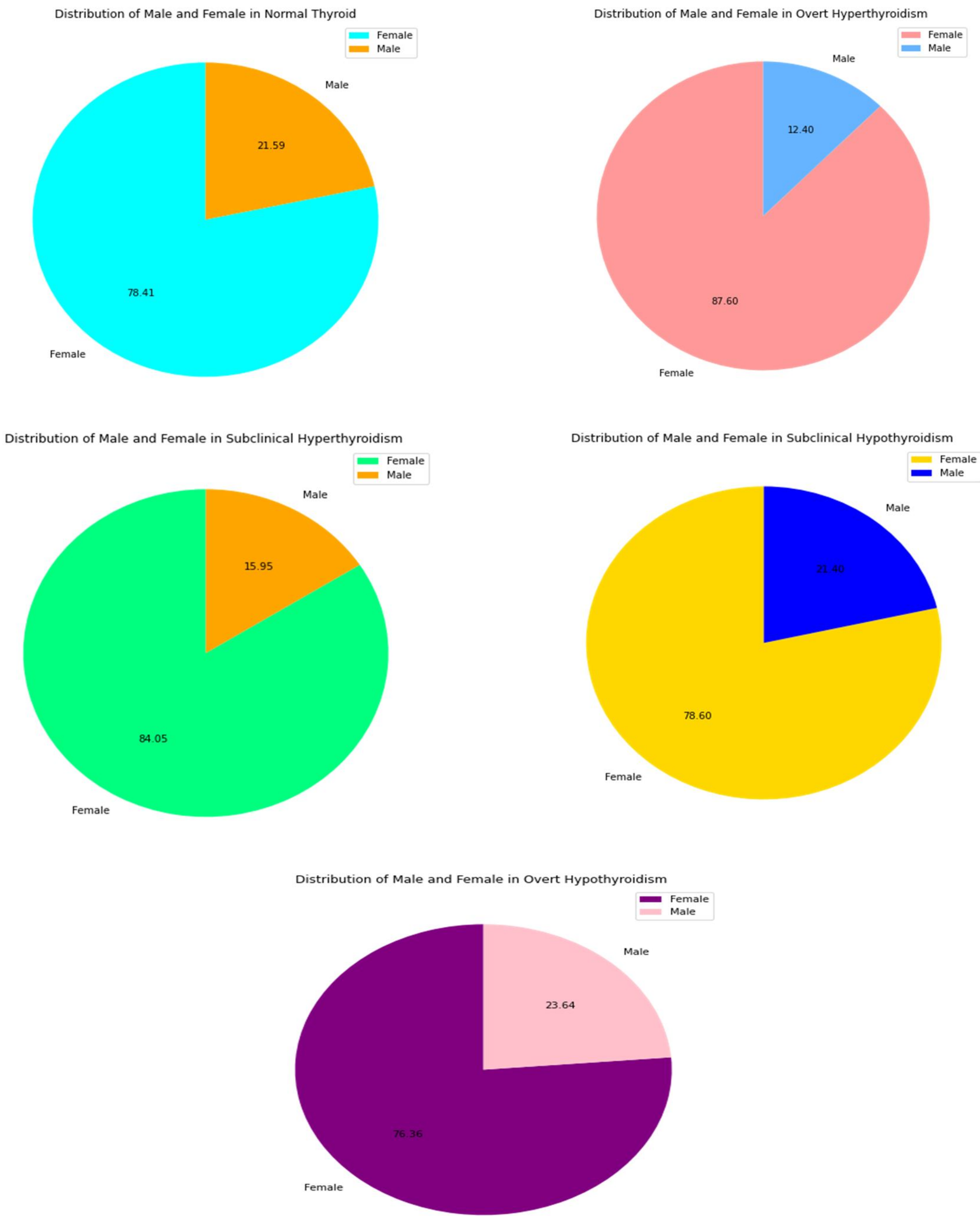


Fig. 6. Frequency Distribution of Thyroid Types and Gender

The tabulated data provides a comprehensive view of gender distributions across distinct categories of thyroid conditions within the dataset. EUTHYROID' group representing individuals with normal thyroid function, there appears to be a substantial imbalance favoring females, comprising 78.41% of this category, while males constitute 21.59%. Overt Hyperthyroidism' demonstrates a considerable female predominance, with 87.6% compared to a smaller percentage of males at 12.4%. Subclinical Hyperthyroidism' displays a distinct trend with 84.05% females and 15.95% males, indicating a substantial female overrepresentation. Subclinical Hypothyroidism' illustrates a relatively higher proportion of females at 78.6%, compared to 21.4% males.

H. Relationship Between T3, T4, Age, And Gender

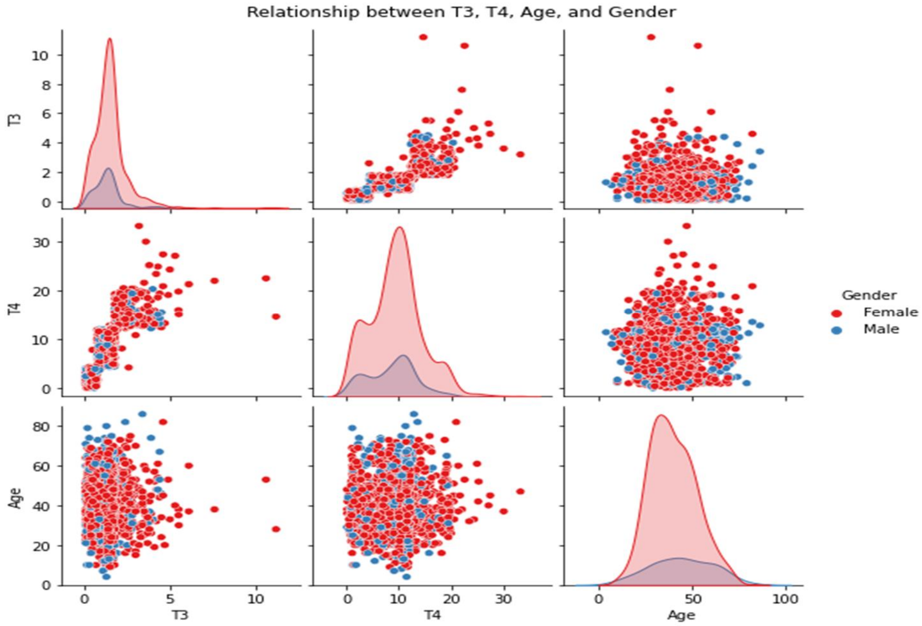


Fig. 7. Relationship Between T3, T4, Age and Gender

By examining these relationships across different variables and genders in the pair plot, researcher can glean insights into potential correlations, patterns, or differences in how T3, T4, and Age relate to each other based on gender. These observations contribute to understanding gender-specific trends or correlations in thyroid hormone levels concerning age in the thyroid dataset.

I. Correlations In Dataset

The statistical measure of correlation quantifies the direction and intensity of the association between two variables. Gaining an understanding of the significance of correlation can yield valuable insights across a multitude of disciplines, encompassing scientific inquiry, data examination, and the formulation of decisions. Understanding and identifying the relationship between two variables is facilitated by correlation. The metric offers a quantitative indication of whether the variables are positively correlated (experience a simultaneous increase or decrease), negatively correlated (one variable increases while the other decreases), or do not demonstrate any statistically significant correlation. High-correlation variables may provide redundant information, while low-correlation variables may have little predictive power.

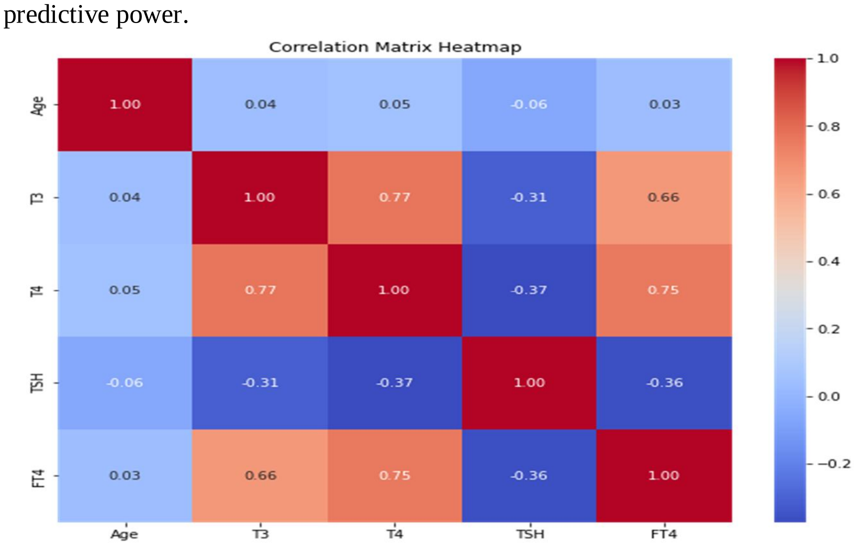


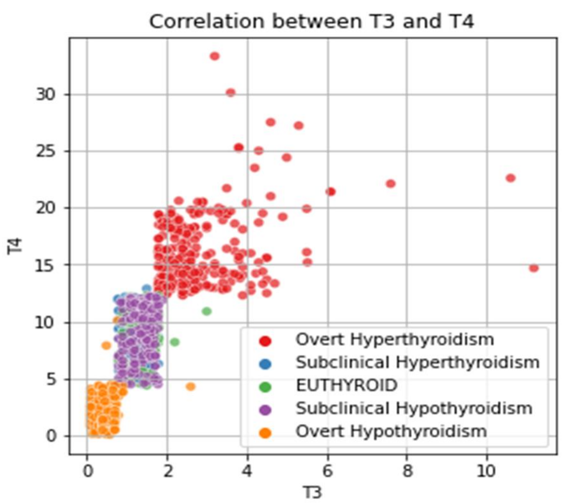
Fig. 8. Correlation Matrix

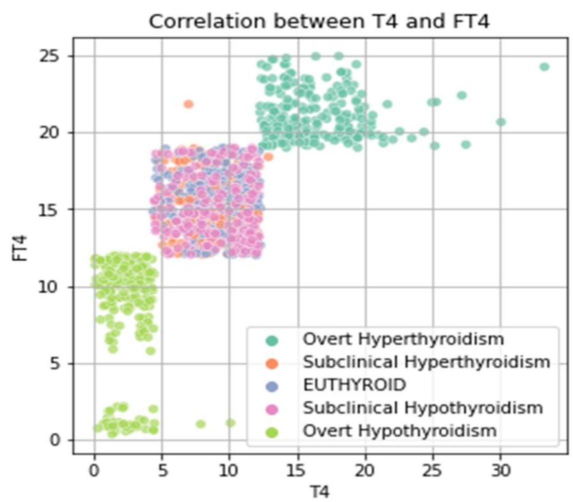
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

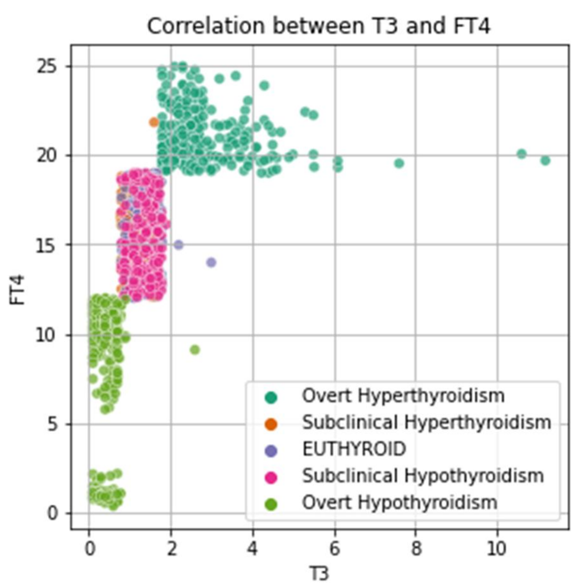
Table 1 Top Correlated Features

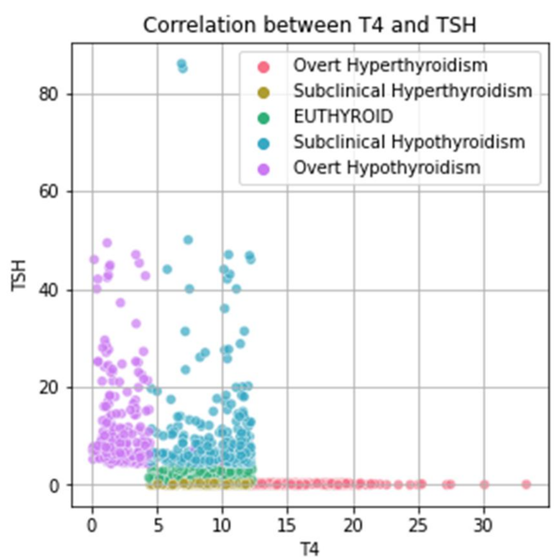
Feature 1	Feature 2	Value
T3	T4	0.765527
T4	FT4	0.754446
T3	FT4	0.658159
T4	TSH	0.373470
TSH	FT4	0.357850
T3	TSH	0.313228

T3 and T4 exhibit a notably strong positive correlation of approximately 0.77. The correlation coefficient of around 0.75 suggests a strong positive relationship between total thyroxine (T4) and free thyroxine (FT4), indicating that variations in one are closely associated with changes in the other. This correlation around 0.66 represents a moderate to strong positive relationship between triiodothyronine (T3) and free thyroxine (FT4), indicating a significant association between these hormone levels. The correlation coefficient of approximately 0.37 suggests a moderate positive relationship between total thyroxine (T4) and thyroid-stimulating hormone (TSH). The correlation of about 0.36 implies a moderate positive association between thyroid-stimulating hormone (TSH) and free thyroxine (FT4). With a correlation coefficient around 0.31, indicating some level of association between triiodothyronine (T3) and thyroid-stimulating hormone (TSH).









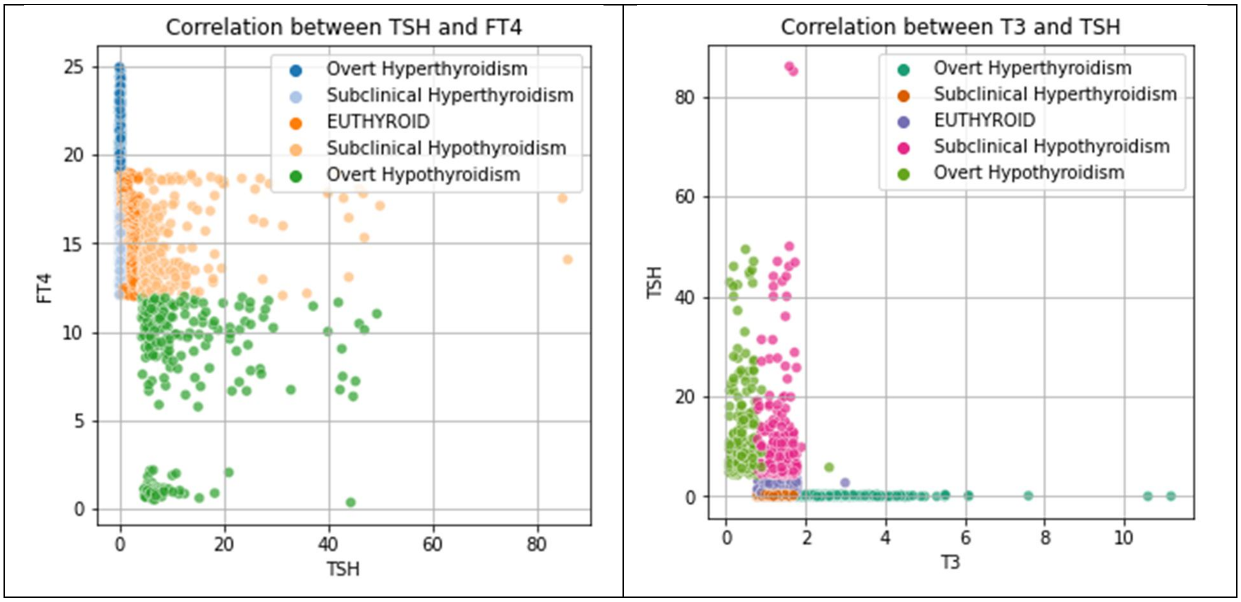


Fig. 9. Correlation Between T3, T4, TSH, FT4

V. CONCLUSION

Since it is the most important thing to keep data set ready for the application of classification algorithm training available irrespective of the size of information and data mining skills, if the researcher does not make sense of data collection, a machine would be almost useless or even harmful. The exploratory analysis of thyroid dataset revealed compelling insights into the relationships between hormone levels, age, and gender. This phase sets the foundation for subsequent analyses and provides valuable insights into thyroid function, paving the way for further investigation and clinical implications. Key insights derived from the EDA is emphasizing gender-based disparities, age-related trends, and interrelationships among thyroid hormones.

REFERENCES

[1] Abed, S., Saji, S., & Alshayeji, M. H., "Predictive Analysis of Early Thyroid Disorders Using Integration of Data Mining and Ensemble Intelligence Approaches", International Journal of Intelligent Systems,2026.

[2] Alam, M. Z., Rahman, R., Sozib, H. M., H. A., A. H., & Sabeena, A. A., “Enhancing Thyroid Disease Diagnosis With Machine Learning and Counterfactual Explainable AI”, IEEE, 2026.

[3] Hassan, A., Ramzan , S., Raza, A., Iqbal, M. M., Smerat , A., Fitriyani, N. L., . . . Lee, S. W. , “Improving thyroid disorder diagnosis via innovative stacking ensemble learning model”, Digit Health, 2025.

[4] Islam , S. S., Haque, M., Miah , M. U., Sarwar, T. B., & Nugraha , R., “Application of machine learning algorithms to predict the thyroid disease risk: an experimental comparative study”, PeerJ Comput Sci, 2022.

[5] Moharekar, D. T., Vadar, M. S., Pol, D. R., Bhaskar, D. C., & Moharekar , M. J., “Thyroid Disease Detection Using Machine Learning and Pycaret”, Specialusis Ugdymas / Special Education, 2022.

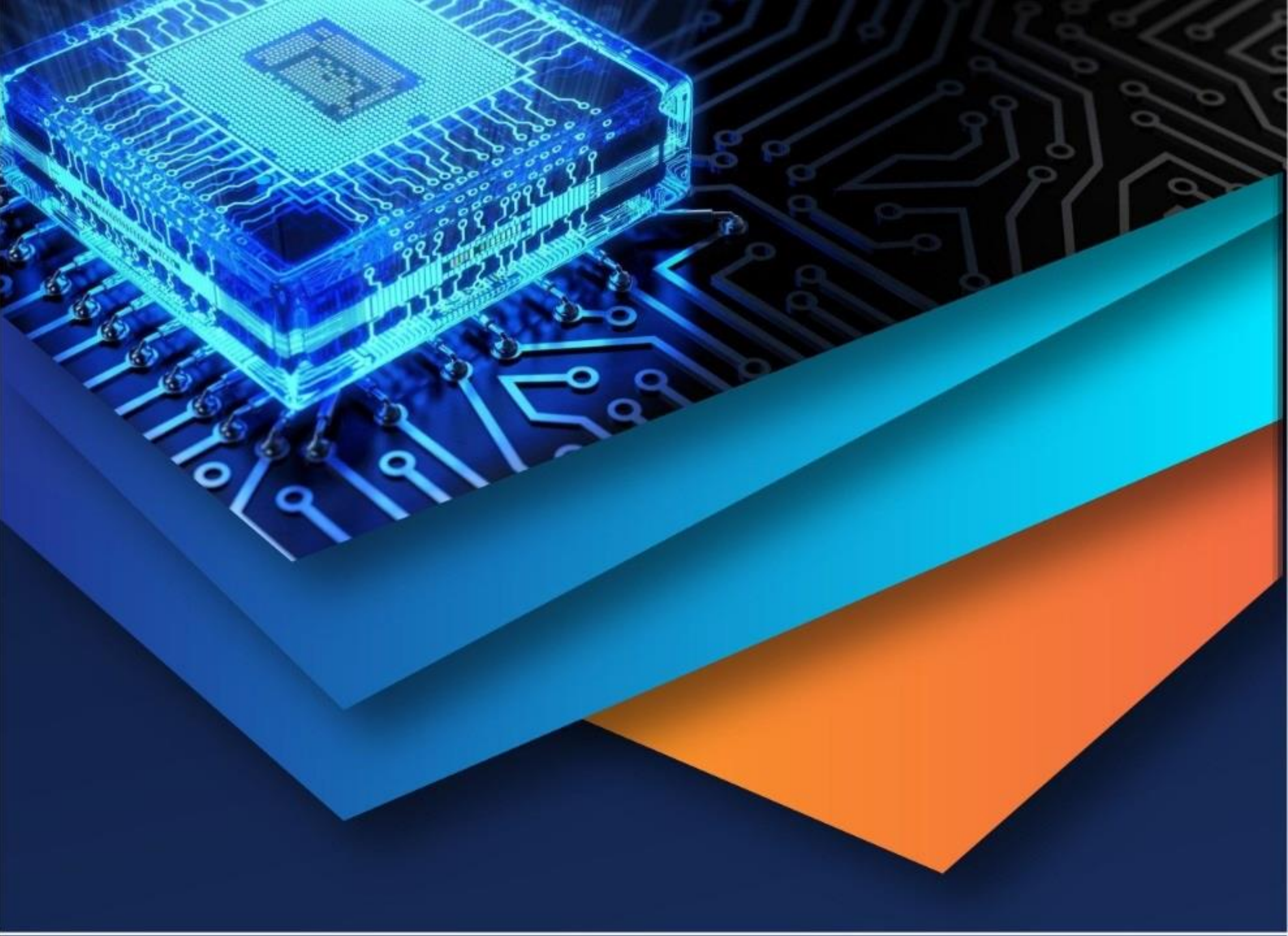
[6] Mollica , G., Francesconi , D., Costante , G., Moretti , S., Giannini, R., Puxeddu, E., & Valigi, P., “Classification of Thyroid Diseases Using Machine Learning and Bayesian Graph Algorithms”, IFAC-PapersOnLine. Elsevier, 2022.

[7] Osasere , O., Iqbal, M. Z., & Xining (Ning) , W., “Enhanced Diagnosis of Thyroid Diseases Through Advanced Machine Learning Methodologies”, Sci, 2025.

[8] S I T A L D I N, N. “Thyroid Disease Prediction and Symptom Exploration During Early Pregnancy”, 2022

[9] Thalor, M., Pathak, M., Kale, V., & Bhende, V., “Automated Diagnosis to Predict the Thyroid Using Machine Learning Algorithms”, SSRG International Journal of Electrical and Electronics Engineering, 2024.

[10] Tiwari, Y., Saxena, D., Pal, P., Dixit, H. M., & Yadav, A., “Thyroid Disease Prediction: Leveraging Machine Learning for Early Recurrence Detection”, *12th International Conference on Computing for Sustainable Global Development (INDIACom), Delhi, India: IEEE,* 2025



10.22214/IJRASET

INDEX COPERNICUS
45.98

IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY