e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 7, July 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 7, July 2024 |

| DOI:10.15680/IJMRSET.2024.0707060 |

Thyroid Disease Detection using Machine Learning

Rani, Sadhana Kumble

PG Student, Department of MCA, Mangalore Institute of Technology & Engineering, Moodarbidri, Karnataka, India

Assistant Professor, Department of MCA, Mangalore Institute of Technology & Engineering, Moodarbidri, Karnataka, India

ABSTRACT: Thyroid disorders have a major influence on public health systems and are becoming a more pressing global health concern. For these illnesses to be effectively treated, accurate prognosis is essential. This research comprehensively assesses the predictive power of five popular machine learning (ML) algorithms: K-Nearest Neighbours (KNN), Decision Trees Classifier (DTC), Random Forest classifier (RFC), and Logistic Regression classifier (LRC) using a sizable dataset.

Simple binary result prediction is where logistic regression shines, but decision trees provide a systematic way to make decisions on a range of tasks. KNN improves the usefulness of new data in pattern recognition by classifying it according to how similar it is to existing data points. To increase accuracy, Random Forest integrates forecasts via several decision trees using ensemble learning. Thorough evaluations are conducted on each algorithm's performance; Decision Trees yielded the best accuracy at 99%, KNN at 93%, and Logistic Regression at 88%. Additionally, Random Forest performed admirably when it came to predicting thyroid problems.

This review highlights the usefulness among various machine learning techniques in healthcare settings, especially regarding enhancing patient care outcomes and diagnostic accuracy, additionally to closely examining their effectiveness. This project is to improve public health management strategies by equipping medical practitioners with trustworthy instruments for early thyroid problem identification and intervention through the advancement of predictive capacities.

KEYWORDS: Machine Learning, Decision Tree, KNN, Logistic Regression, Random Forest, Naive Bayes, Thyroid Disease Prediction.

I.INTRODUCTION

Thyroid diseases represent a significant global health concern, affecting millions worldwide, particularly prevalent among women aged 17 to 54. In India, approximately one in ten individuals suffers from thyroid disease, underscoring its widespread impact. Similarly, in Bangladesh, the condition affects about 50 million people, predominantly women, yet awareness remains inadequate in many regions.Located in the neck, the thyroid gland produces essential hormones T4 (thyroxine) and T3 (triiodothyronine), crucial for regulating metabolism, body temperature, and cellular function. Dysfunctions in hormone production lead to three main categories of thyroid conditions: Eu-Thyroidism (normal hormone levels), Hyper-Thyroidism (excessive hormone levels), and Hypo-Thyroidism (insufficient hormone levels).

Low T3 and T4 levels and increased TSH (thyroid-stimulating hormone) are the hallmarks of hypothyroidism, which can lead to symptoms including sadness, weight gain, and exhaustion. Conversely, hyperthyroidism is characterized by low TSH levels along with high T3 and T4 levels, which causes symptoms including anxiety, weight loss, and an elevated heart rate. If left untreated, these illnesses might result in serious side effects such as heart problems, infertility, and psychological difficulties. Given these ramifications, successful patient management and treatment depend on early identification and accurate forecasting. Effective as they are, traditional diagnostic techniques frequently need invasive procedures, which highlights. The requirement for effective non-invasive prediction tools to support early detection. Professionals in healthcare today possess sophisticated tools at their disposal thanks to the field of machine learning (ML) techniques, which can handle huge datasets and enhance the classification as well as forecasting illness. In this investigation, we assess the predictive power of five machine learning algorithms: K-Nearest Neighbors (KNN), Random Forest, Decision Trees, and Logistic Regression.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 7, July 2024 |

| DOI:10.15680/IJMRSET.2024.0707060 |

Binary outcome prediction is made easier by logistic regression, which can differentiate between thyroid illness and not. By organizing decisions in a hierarchical manner, decision trees offer models that are simple to comprehend and provide for lucid comprehension of the decision-making process. When classifying incoming data points, KNN uses similarity to classify existing data, this is effective in identifying patterns. To improve accuracy, an ensemble method known as Random Forest aggregates the outcomes of many decision trees.

We do thorough pre-processing on UC Irvine Machine Understanding Repository data and assess each method's predicting ability for thyroid disease. Decision Trees beat all other models with an accuracy rate of 99%, the function of logistic regression is ranked second with 96% accuracy, while Logistic Regression is ranked third with 92%. These findings demonstrate the potential application concerning machine learning (ML) to improve thyroid disease prediction models, providing physicians with more data to work with and improved patient outcomes.

This study offers a thorough review of thyroid illness machine learning-based prediction. The following sections make up its division: Introduction, Related Works, Proposed Methodology, Results and Discussion, and Conclusion. With the intent of improving patient care and health outcomes through better prediction and management techniques, it offers insights into useful applications.

II.RELATED WORK

1. In [1] Authors investigate feature selection and machine learning strategies for forecasting thyroid disorders, concentrating on the categorization of hyperthyroidism and hypothyroidism. They achieve 92.92% accuracy by using Support Vector Machines and Recursive Feature Elimination. Important factors in the diagnosing process include age, sex, and thyroid-related hormone levels (TSH, TT4, T4U, T3, FTI). Their method helps medical decision-making by helping to categorize thyroid illnesses into four groups.

2. By taking into account 10 illness kinds, the research in [2] addresses class imbalance and binary classification in the identification of thyroid disease. It uses differential evolution (DE) to enhance machine learning models, and using an AdaBoost model that has been tuned, it achieves an accuracy of 0.998. For data augmentation, conditional generative adversarial networks (CTGAN) are utilized to lessen bias and enhance model performance. The optimized models produce reliable and broadly applicable findings, and ensemble models deliver superior results compared to linear ones.

3. The authors of [3] suggest predicting thyroid illness using K-Nearest Neighbors (KNN), Decision Tree (DT), and Models of Multi-Layer Perceptrons (MLP). MLP yields a precision of 95.73% and an Area Under the Curve (AUC) of 94.23%. Using a dataset of 3163 patients and 24 thyroid features, the study analyzes different models and finds that MLP performs better than KNN and DT. They emphasize the relevance of machine learning in medical diagnostics by highlighting MLP's potential for accurate categorization of thyroid illnesses and suggesting its application to other chronic ailments including cancer and heart disease.

4. Utilizing classifiers such as Support Vector Machine (SVM) and Decision Tree (DT) and feature selection techniques such as Recursive Feature Selection (RFE), Random Forest (RF), Logistic Regression (LR), and Naive Bayes (NB), [4] focuses on forecasting hypothyroid illness. They intend to improve prediction accuracy and boost public health efforts by working with larger datasets in the future. They now obtain a constant 99.35% accuracy with RFE across all classifiers.

5. The research study investigates the application of Featurewiz, a library with feature selection, and machine learning to predict thyroid illness. A dataset of Australian thyroid patients accustomed to train many models, including Decision Tree, K-Nearest Neighbor, Support Vector Classifier, Naïve Bayes, and Logistic Regression, Random Forest, and XGBoost. Based on evaluations, Random Forest with Featurewiz fared superior than other models, with an accuracy rate of 99.45%. This study demonstrates the benefits of ensemble techniques for accurate prediction of thyroid illness and the efficacy of Featurewiz in feature selection.

6. Using feature selection as well as machine learning approaches, the author of [5] investigates the prediction of thyroid illness and achieves a high accuracy of 0.99 with random forest. Their work emphasizes the superiority preference for machine learning over deep learning in this field and emphasizes the necessity of more extensive class classification and bigger datasets for further advancements.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 7, July 2024 |

| DOI:10.15680/IJMRSET.2024.0707060 |

III.METHODOLOGY

On this topic venture, we utilized systematic machine learning methods to predict thyroid diseases. Data preparation, feature selection, model training, data collection, and assessment were the main components of the technique. The functioning flow of the recommended system is shown in the diagram above. Its objective is to develop a robust forecasting model for thyroid disease detection. To accomplish this, data preparation, including data cleansing and validation, is necessary. Following that, the information it requests to be updated. Since data augmentation utilizes pre-existing data to create modified copies of datasets, therefore artificially expanding the instruction set. After data augmentation, the data must be educated using a range of Decision trees are examples of machine learning techniques, logistic regression, K-nearest neighbors, and the Random Forest classifier approach. Our research on using methods from machine learning to the prediction of thyroid illness is founded on this dataset.

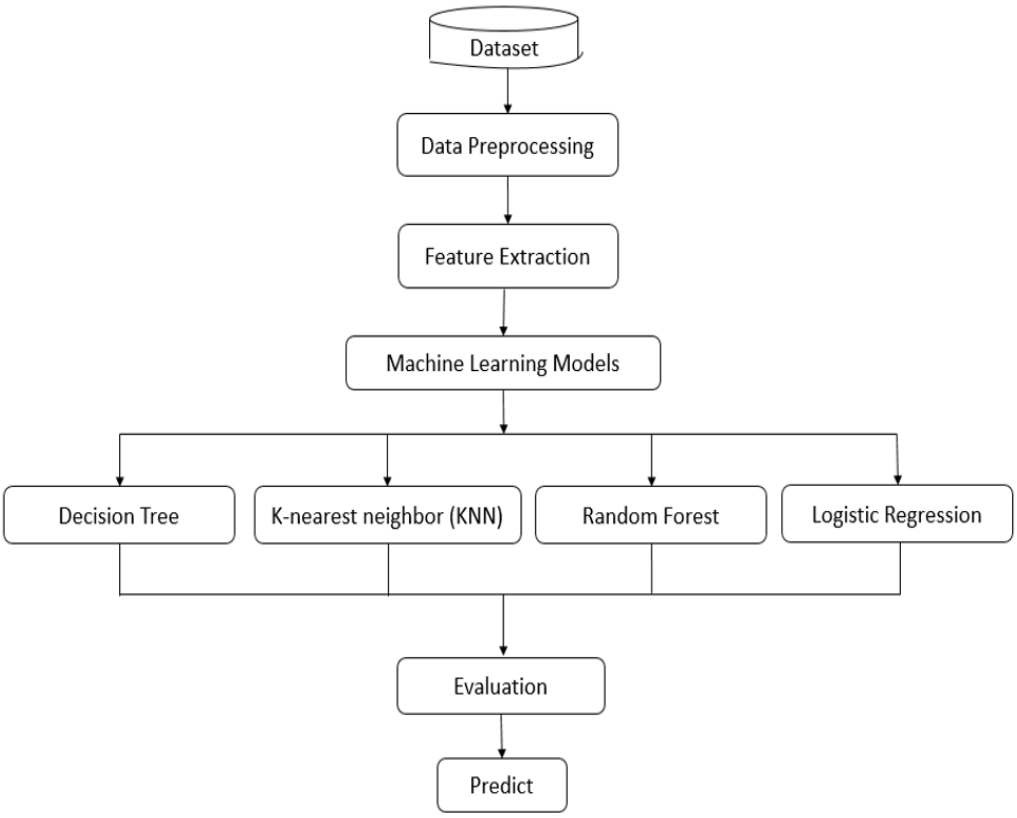

Fig. 1 System Dataflow Diagram

A. Data Collection

In the present investigation, we used a Kaggle dataset that included 2667 examples with 29 features. Thyroid illness outcomes are categorized in this dataset into four classes: primary hypothyroid, secondary hypothyroid, negative, and compensated hypothyroid. It was used to train algorithms that were supposed to forecast thyroid illness based on test findings and patient symptoms. Robust model training, which is necessary for precise illness categorization and prediction in clinical applications, was made possible by the dataset's abundant occurrence count and broad attribute collection.

B. Data Preprocessing

An essential first step in improving the quality of data for analysis is data preparation. Originally, We eliminated from our dataset every columns that had 100% missing values. The median was used to impute missing values for numerical data, while the most common values were used to fill in the gaps for categorical variables. After that, outliers were eliminated to guarantee the accuracy of the information and the stability of our models. To ensure that improve model compatibility, categorical characteristics were encoded using 'OrdinalEncoder' and converted into numeric

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 7, July 2024 |

| DOI:10.15680/IJMRSET.2024.0707060 |

representations. 'StandardScaler' accustomed to standardize the range of numerical characteristics to be able to facilitate efficient model training. Last but not least, the dataset was balanced using the Synthetic Minority Over-sampling Technique (SMOTE) to resolve the class imbalance. This made certain that every class was adequately represented for precise predictive modelling and analysis.

C. Feature Selection

A methodical procedure to find the most pertinent features for our model is to use 'SequentialFeatureSelector' for forward feature selection. It starts with an empty set and adds features repeatedly until the model performs better than the baseline, usually as shown by metrics like accuracy or another predetermined criterion. This iterative process keeps going until a predetermined number of features are reached or performance can no longer be improved upon. SequentialFeatureSelector optimizes overall model performance and minimizes over fitting by concentrating on the subset of features that contribute most to predictive power through a systematic evaluation of feature subsets. This improves interpretability, efficiency, and generalization ability of the model.

D. Model Selection and Training

This stage involves applying several machine learning algorithms to the preprocessed data in order to identify the most accurate model for thyroid illness prediction. The primary algorithms in use are:

- K-Nearest Neighbors (KNN): KNN is a classification technique that uses the majority class among its k-nearest neighbors in the training data to predict the class of a new instance. To assess how similar two instances are, it uses distance measures like the Euclidean distance.
- Logistic Regression: Using a logistic function, logistic regression estimates the likelihood of a binary outcome. It is especially well-suited for jobs involving binary classification, such as determining the probability of thyroid illness existence or absence.
- Decision Trees: Based on feature values, Decision Trees provide a model of decisions that resembles a tree. By dividing the data at each node according to the most illuminating aspect, they created a model that is simple to understand and display.
- Random Forest Classifier (RFC): Using a random subset of the data for training, RFC creates an ensemble of decision trees. The final categorization is determined by adding the predictions from each tree. This method reduces overfitting and increases precision.

Eighty percent of the dataset is utilized for training and twenty percent is put aside for testing. The dataset is divided into training and testing sets. In order to generate predictions, the models must first identify patterns in the training set of data. To discover the best parameters for each model, hyperparameter tuning is done using methods like Grid Search CV and Randomized Search CV. While Randomized Search CV selects samples from a range of potential hyperparameters, Grid Search CV thoroughly investigates a range of hyperparameters.

Cross-Validation: To make sure that the model performs consistently across several training data subsets, K-fold cross-validation is employed. Using this approach, the data is divided into k subsets. The model is then trained on k-1 subsets, and its validity is checked on the remaining subset. In order to evaluate the model's capacity for generalization, the procedure is repeated k times, using one subset as the validation set each time. The outcomes are then averaged.

E. Model Evaluation

During the model assessment phase, several algorithms' performances are evaluated to discover which one is better for predicting thyroid sickness. This stage consists of several important parts. First, the models are assessed using performance indicators including recall, accuracy, precision, and F1-score. The model's overall correctness is measured by accuracy, and its capacity to discriminate between positive and negative instances is shown by precision and recall. The F1-score strikes a balance between recall and accuracy to offer a single indicator of model performance. Second, the performance indicators are illustrated using visualization approaches including confusion matrices that show the ratios of true positives, true negatives, false positives, and false negatives, as well as bar charts for recall, accuracy, precision, and F1-score. These illustrations aid in highlighting the advantages and disadvantages of each paradigm.

Finally, the best-performing model is chosen in accordance with the assessment results. The model that has the best balance between accuracy, precision, recall, and F1-score is the one that is selected, guaranteeing that the model of choice is the most appropriate for forecasting thyroid ailment. This thorough evaluation process ensures that the chosen model produces accurate and dependable forecasts of thyroid ailment, qualifying it for application in clinical settings.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 7, July 2024 |

| DOI:10.15680/IJMRSET.2024.0707060 |

F. Feature Analysis and Prediction

Understanding the significance of different characteristics in predicting thyroid ailment is the goal of feature analysis. In order to discover important attributes, this stage entails assessing the impact of various characteristics using techniques like feature significance ranking. It is possible to create more effective treatment and diagnostic plans by comprehending these essential components. The trained model then applies this understanding to provide predictions in real-time based on fresh patient data, providing insightful information about disease trends that might enhance treatment strategies and patient care.

IV. PROPOSED SYSTEM

Using a Random Forest classifier trained on a CSV dataset comprising vital patient data like thyroid hormone levels, symptoms, and demographics, the suggested approach seeks to produce a trustworthy model for the prediction of thyroid illness. In comparison to individual trees, this ensemble learning approach builds several decision trees during training and aggregates their predictions to increase accuracy and decrease overfitting. Metrics such as F1-score, recall, accuracy, and precision shall be applied in the model's performance evaluation. These metrics are essential to ascertain the model's capacity to accurately identify patients with thyroid illness while reducing false negatives as well as false positives. The F1-score offers a comprehensive evaluation of the efficacy of the model in clinical diagnostic situations by striking a balance between precision and recall.

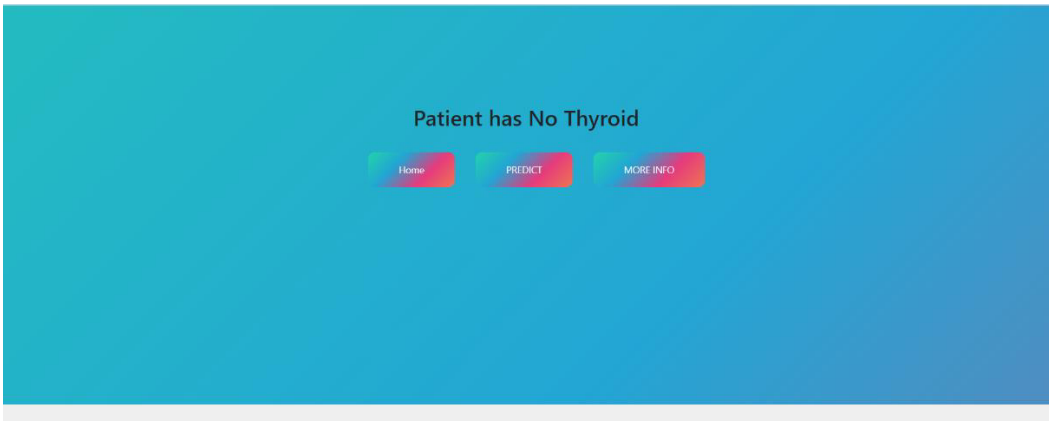

Fig. 2 Output of the Proposed System

Making Use of cross-validation techniques, which validate the model's performance across several dataset subsets, will guarantee the model's resilience and generalizability. Iteratively replicating real-world events and upgrading the model derived from fresh data insights improves dependability.

Additionally, feature significance analysis will be employed to ascertain the patient data points that have the most predictive power for thyroid disease. Healthcare professionals can get important insights from this study that will help them better understand illness patterns and develop treatment plans. These elements are perfectly integrated by the Flask application interface, which lets users enter their information for a real-time likelihood prediction of thyroid illness. The system intends to improve healthcare outcomes through early detection and individualized intervention by merging powerful machine learning with user-centric design. This demonstrates the transformational potential of technology in medical diagnosis and patient care.

V. FUTURE WORK

As an extension of the Random Forest model, future research on the thyroid illness detection system may involve investigating more sophisticated machine learning methods like neural networks in an effort to identify more complex associations in thyroid disease data. To enhance the model's prediction ability and resilience, more extensive datasets that include genetic, environmental, and lifestyle components should be included. Enhancing user engagement with intuitive visuals and individualized health insights obtained from the model's predictions might be one way to improve the Flask application interface. Better user participation and adherence to health monitoring procedures may result from this. To guarantee that the system stays adaptable and efficient in changing clinical situations, continuous model

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 7, July 2024 |

| DOI:10.15680/IJMRSET.2024.0707060 |

refinement through continual learning strategies—such as real-time updates with new data and feedback loops from users and healthcare professionals—is necessary.

Finally, to confirm the system's accuracy, dependability, and usefulness in actual clinical settings, it would be crucial to carry out thorough validation studies in association with healthcare practitioners. By using an iterative method, the system would be more prepared for general use, which would ultimately improve patient care outcomes and diagnostic accuracy for thyroid disease oversight.

VI.CONCLUSION

To determine which patient the most crucial components are the data points. in forecasting thyroid illness, the system would furthermore comprise feature significance analysis. Healthcare practitioners can benefit greatly from this analysis's insightful observations, this may be beneficial them better understand illness patterns and develop more specialized and efficient treatment programs. Through the incorporation of these components, the proposed framework seeks to offer a dependable and effective instrument for the identification of thyroid illness, eventually enhancing patient outcomes via prompt diagnosis and treatment.

This iterative approach keeps the model current and accurate, transforming it into a vital tool for the early diagnosis and treatment of thyroid problems. To determine which patient the most crucial components are the data points in forecasting thyroid illness, the system will furthermore have a feature significance analysis. Healthcare practitioners can benefit greatly from this analysis's insightful observations, it can aid in their creation of additional customized and effective treatment plans by helping them comprehend sickness patterns. Through the combination of these elements, the suggested system seeks to offer a dependable and effective instrument for the identification of thyroid illness, eventually enhancing patient outcomes via prompt diagnosis and treatment.

REFERENCES

[1] P. Duggal and S. Shukla, "Prediction Of Thyroid Disorders Using Advanced Machine Learning Techniques," 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2020, pp. 670-675, doi: 10.1109/Confluence47617.2020.9058102.

[2] Gupta, P., Rustam, F., Kanwal, K. et al. Detecting Thyroid Disease Using Optimized Machine Learning Model Based on Differential Evolution. Int J Comput Intell Syst 17, 3 (2024). https://doi.org/10.1007/s44196-023-00388-2

[3] M. Pal, S. Parija and G. Panda, "Enhanced Prediction of Thyroid Disease Using Machine Learning Method," 2022 IEEE VLSI Device Circuit and System (VLSI DCS), Kolkata, India, 2022, pp. 199-204, doi: 10.1109/VLSIDCS53788.2022.9811472.

[4] M. Riajuliislam, K. Z. Rahim and A. Mahmud, "Prediction of Thyroid Disease(Hypothyroid) in Early Stage Using Feature Selection and Classification Techniques," 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), Dhaka, Bangladesh, 2021, pp. 60-64, doi: 10.1109/ICICT4SD50815.2021.9397052.

[5] Chaganti R, Rustam F, De La Torre Díez I, Mazón JLV, Rodríguez CL, Ashraf I. Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques. Cancers (Basel). 2022 Aug 13;14(16):3914. doi: 10.3390/cancers14163914. PMID: 36010907; PMCID: PMC9405591.

[6] R. Rao and B. S. Renuka, "A Machine Learning Approach to Predict Thyroid Disease at Early Stages of Diagnosis," 2020 IEEE International Conference for Innovation in Technology (INOCON), Bangluru, India, 2020, pp. 1-4, doi: 10.1109/INOCON50539.2020.9298252.

[7] Islam SS, Haque MS, Miah MSU, Sarwar TB, Nugraha R. Application of machine learning algorithms to predict the thyroid disease risk: an experimental comparative study. PeerJ Comput Sci. 2022 Mar 3;8:e898. doi: 10.7717/peerj-cs.898.

[8] Begum, Amina, and A. Parkavi. "Prediction of thyroid disease using data mining techniques." 2019 5th international conference on advanced computing & communication systems (ICACCS). IEEE, 2019.

[9] Tyagi, R. Mehra and A. Saxena, "Interactive Thyroid Disease Prediction System Using Machine Learning Technique," 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan, India, 2018, pp. 689-693, doi: 10.1109/PDGC.2018.8745910.

[10] Mir, Yasir Iqbal, and Sonu Mittal. "Thyroid disease prediction using hybrid machine learning techniques: An effective framework." International Journal of Scientific & Technology Research 9.2 (2020): 2868-2874.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 7, July 2024 |

| DOI:10.15680/IJMRSET.2024.0707060 |

[11] Raisinghani, Sagar, et al. "Thyroid prediction using machine learning techniques." Advances in Computing and Data Sciences: Third International Conference, ICACDS 2019, Ghaziabad, India, April 12–13, 2019, Revised Selected Papers, Part I 3. Springer Singapore, 2019.

[12] Z. J. Peya, M. Shymon Islam and M. K. Naher Chumki, "Thyroid Disease Prediction based on Feature Selection and Machine Learning," 2022 25th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 2022, pp. 495-500, doi: 10.1109/ICCIT57492.2022.10054746.

[13] V. Vinodhini and F. K, "Prediction of Thyroid Disease Using Machine Learning Algorithms," 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2023, pp. 2813-2817, doi: 10.1109/ICACITE57410.2023.10183108.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com