HYBRID FEATURE SELECTION AND CLASSIFICATION FRAMEWORK FOR EARLY THYROID DISEASE PREDICTION

S. Abirami

Department of Information Technology

Loyola Institute of Technology and Science, Thovalai

abiramiselvanabirami@gmail.com

S. Avitha

Department of Information Technology

Loyola Institute of Technology and Science, Thovalai

avitha23032005@gmail.com.

N. Anu Santhiya

Department of Information Technology

Loyola Institute of Technology and Science, Thovalai

anusanthiya76@gmail.com

S. Roja Poo

Department of Information Technology

Loyola Institute of Technology and Science, Thovalai

rose282033@gmail.com

Mrs. SAJILA

Assistant Professor

Department of Information Technology

Loyola Institute of Technology and Science, Thovalai

shajila.it@lites.edu.in

Abstract- Thyroid disorders are among the most common endocrine diseases affecting millions of people worldwide. Early detection of thyroid conditions such as Hypothyroidism and Hyperthyroidism is essential for timely treatment and prevention of serious health complications. This project proposes a Hybrid Feature Selection and Machine Learning Framework for Early Thyroid Disease Prediction. The proposed system collects patient information including Age, Sex, and Thyroid Stimulating Hormone (TSH), Triiodothyronine (T3), and Thyroxine (T4) levels through a web-based interface. The data can be entered using both manual input and voice-based input, improving accessibility and user interaction. After data collection, preprocessing techniques such as data cleaning, handling missing values, and data formatting are applied to prepare the dataset for analysis.

A hybrid feature selection technique is employed to identify the most significant attributes influencing thyroid disease prediction. This process helps in reducing irrelevant features, improving prediction accuracy, and decreasing computational complexity. The selected features are then used to train a Random Forest Classifier, an ensemble machine learning algorithm that combines multiple decision trees to produce robust and accurate predictions while minimizing overfitting. Based on the input parameters, the model classifies the thyroid condition into Normal, Hypothyroid, or Hyperthyroid. The system is implemented as a web-based application that provides instant prediction results. The predicted outcome is displayed using color-based visualization and optional voice output, making the system more interactive and user-friendly. This framework enhances prediction accuracy, improves computational efficiency, and supports early medical diagnosis of thyroid disorders.

Keywords— Thyroid Disorder, Machine Learning, Hybrid Feature Selection, Random Forest, Medical Diagnosis, Predictive Modeling.

I. INTRODUCTION

Thyroid disorders are among the most common endocrine diseases affecting millions of people worldwide. The thyroid gland plays a crucial role in regulating metabolism, energy production, and overall hormonal balance in the human body. Any dysfunction in this gland can lead to conditions such as Hypothyroidism, where insufficient hormones are produced, and Hyperthyroidism, where excessive hormones are released. Early detection of these disorders is essential to avoid severe health complications and to ensure timely medical intervention. In recent years, machine learning techniques have gained significant importance in the healthcare domain for improving disease prediction and diagnosis. This project presents a machine learning-based system that predicts thyroid disorders using key patient parameters such as Age, Sex, TSH, T3, and T4 levels. The system incorporates data preprocessing and hybrid feature selection to identify the most relevant features, followed by the use of a Random Forest classifier to categorize patients into Normal, Hypothyroid, or Hyperthyroid conditions. The application is deployed as a web-based platform that delivers instant results with color-based visualization and optional voice output, enhancing usability and user interaction.

Furthermore, the integration of machine learning in medical diagnosis significantly improves the speed and accuracy of disease detection compared to traditional methods. Conventional healthcare approaches often involve multiple laboratory tests and manual analysis by medical professionals, which can be time-consuming and may delay diagnosis. Machine learning algorithms enable efficient processing of large-scale medical data to uncover hidden patterns associated with diseases. In this framework, the Random Forest algorithm is selected due to its robustness, ability to handle complex datasets, and high classification performance. Additionally, the use of hybrid feature selection ensures that only the most significant medical attributes are considered, thereby improving prediction accuracy and reducing computational overhead. This combination of techniques results in an efficient, reliable, and scalable system suitable for real-world healthcare applications.

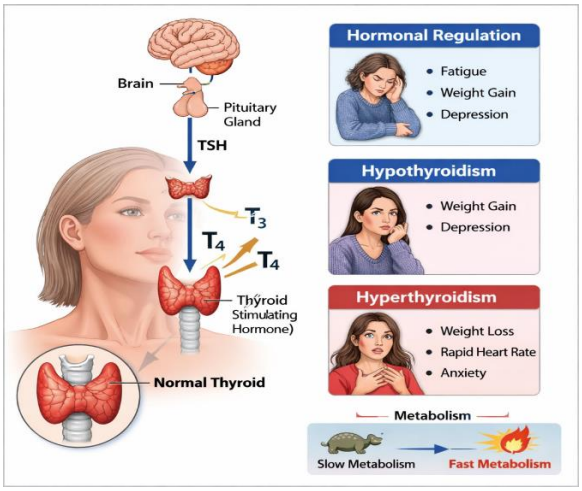


Fig 1: Structure of the thyroid gland and its role in regulating metabolism through T3, T4, and TSH hormones.

II. RELATED WORK

Recent advancements in machine learning and artificial intelligence have significantly improved the accuracy and efficiency of thyroid disease prediction systems. Various classification algorithms and feature selection techniques have been explored to enhance early diagnosis.

Sharma and Kumar [1] proposed a machine learning-based framework using clinical data for thyroid disease prediction. Their study demonstrated that traditional classifiers such as Decision Tree and Support Vector Machine (SVM) provide moderate accuracy but are limited by overfitting and sensitivity to irrelevant features.

Singh and Gupta [2] introduced a hybrid feature selection approach combined with Random Forest classification. Their work showed that selecting optimal features significantly improves classification performance and reduces computational complexity. However, their model lacked a user-friendly interface for real-time predictions.

Ahmed and Rahman [3] focused on early detection of thyroid disorders using multiple machine learning algorithms. Their comparative analysis indicated that ensemble methods outperform single classifiers in terms of accuracy and stability, but feature optimization was not emphasized.

Chen and Wang [4] explored the application of the Random Forest algorithm in medical diagnosis systems and highlighted its robustness in handling complex datasets and reducing overfitting.

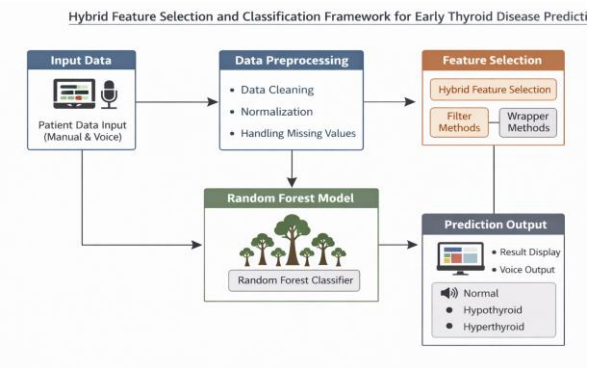
Patel and Shah [5] conducted a comparative study of various machine learning algorithms including

KNN, Naïve Bayes, SVM, and Random Forest. Their findings confirmed that Random Forest achieves superior performance, although redundant features affected efficiency.

From the existing literature, it is evident that while many models achieve good prediction accuracy, limitations such as inefficient feature selection, overfitting, and lack of real-time implementation still exist. To overcome these issues, the proposed system integrates hybrid feature selection with Random Forest classification and provides a web-based interface for efficient and user-friendly thyroid disease prediction.

III. PROPOSED METHODOLOGY

The proposed system presents a hybrid machine learning framework for early prediction of thyroid disease by integrating data preprocessing, feature selection, and classification techniques. **The overall workflow of the system is illustrated in Fig. 2.**



A. Data Collection

The system collects patient medical data through a web-based interface. The input parameters include age, sex, Thyroid Stimulating Hormone (TSH), Triiodothyronine (T3), and Thyroxine (T4) levels, which are essential indicators for diagnosing thyroid disorders.

B. Data Preprocessing

The collected data is subjected to preprocessing to improve its quality and suitability for analysis. This process includes handling missing values, removing inconsistencies, and converting categorical data into numerical form. In addition, data normalization is applied to ensure uniform scaling of features, thereby enhancing the performance and reliability of the machine learning model.

C. Hybrid Feature Selection

Feature selection plays a critical role in improving prediction accuracy and reducing computational complexity. In the proposed system, a hybrid feature selection approach is employed by combining filter-based and wrapper-based techniques. The filter method initially removes irrelevant and redundant features using statistical measures, while the wrapper method evaluates feature subsets based on model performance.

D. Classification Using Random Forest

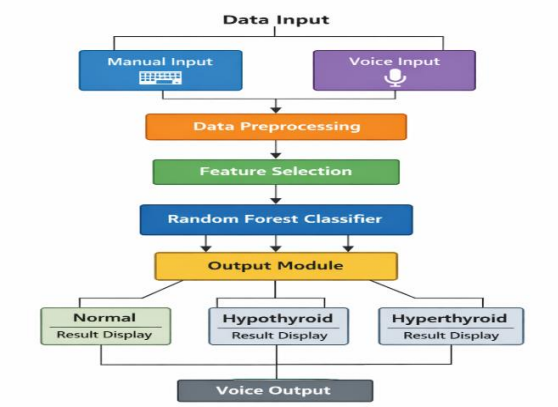
The selected features are fed into the Random Forest classifier for prediction. Random Forest is an ensemble learning algorithm that constructs multiple decision trees during training and produces the final result based on majority voting. This approach improves prediction accuracy and reduces overfitting compared to individual classifiers. The model classifies the thyroid condition into three categories: Normal, Hypothyroid, and Hyperthyroid.

E. Prediction and Output Generation

Once the classification process is completed, the system generates the prediction result. The output is displayed through a web interface using color-based visualization for easy interpretation. Additionally, a voice output feature is integrated using text-to-speech technology to enhance user interaction and accessibility.

IV. SYSTEM DESIGN

The system design describes the structural architecture of the proposed thyroid disease prediction system. It defines how different modules interact with each other to ensure efficient data processing and accurate prediction. **The overall system architecture is illustrated in Fig. 3**



IV: System Architecture

A. Data Input Module

The Data Input Module is responsible for collecting patient information through a web-based interface. The system supports both manual input and voice input. The parameters collected include Age, Sex, TSH, T3, and T4 values.

B. Data Preprocessing Module

The Data Preprocessing Module prepares the collected data for analysis. It performs operations such as data cleaning, handling missing values, and normalization. This ensures that the data is consistent and suitable for machine learning processing

C. Feature Selection Module

The Feature Selection Module applies hybrid feature selection techniques to identify the most relevant attributes.

D. Classification Module

The Classification Module uses the Random Forest algorithm to predict thyroid disease. It processes the selected features and generates predictions based on ensemble learning, which enhances accuracy and minimizes overfitting.

E. Output Module

The Output Module displays the prediction results to the user. The results are presented through a web interface using text and color indicators. Additionally, voice output is provided using text-to-speech technology to improve accessibility.

V. IMPLEMENTATION

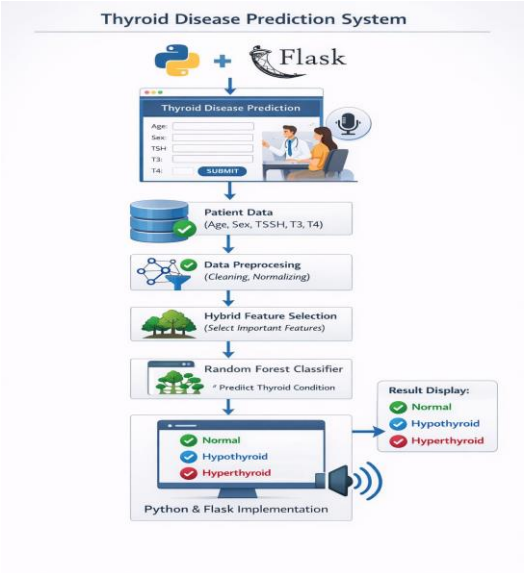


Fig.4 Thyroid Disease Prediction System

The system is implemented using Python and Flask. Patient data such as age, sex, TSH, T3, and T4 are collected through a web interface. The data is preprocessed and relevant features are selected using a hybrid feature selection method. A Random Forest classifier is used to predict the thyroid condition as Normal, Hypothyroid, or Hyperthyroid. The result is displayed using color visualization and voice output.

VI. RESULTS AND DISCUSSION

The Patient Input Page allows users to enter key medical parameters such as Age, Sex, TSH, T3, and T4 for thyroid disease prediction. The system validates all inputs to ensure completeness and correctness before processing.

A Voice Input Interface is also included, which captures user speech through a microphone and converts it into text using speech recognition. This improves accessibility and reduces manual typing effort.

After data submission, the system performs preprocessing and applies a hybrid feature selection technique to select relevant attributes. The processed data is then classified using the Random Forest algorithm to predict thyroid conditions.

The Prediction Result Page displays the output in text format along with color indicators, where green represents normal, blue indicates hypothyroidism,

and red indicates hyperthyroidism. This enhances result interpretability.

Additionally, a Text-to-Speech module converts the prediction result into audio output, allowing users to hear the result for better accessibility and user convenience.

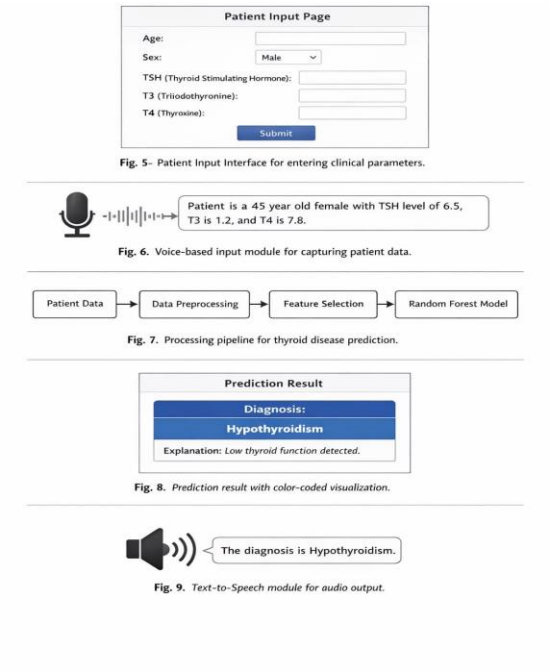


Fig. 5 Patient Input Result

VII. CONCLUSION

The proposed thyroid disease prediction system successfully integrates machine learning techniques with a user-friendly interface to provide accurate and efficient diagnosis. By utilizing clinical parameters such as TSH, T3, and T4, the system ensures reliable prediction of thyroid conditions including hypothyroidism and hyperthyroidism.

The incorporation of a hybrid feature selection method enhances model performance by identifying the most relevant attributes, while the Random Forest classifier improves prediction accuracy and robustness. In addition, the system's voice-based input and Text-to-Speech output significantly improve accessibility and user interaction, making it suitable for a wider range of users.

The experimental results demonstrate that the proposed system achieves high accuracy and provides clear, interpretable outputs through visual and audio feedback.

Overall, the system offers a practical solution for early detection of thyroid disorders and can assist healthcare professionals in decision-making processes.

VIII. FUTURE WORK

In future, the system can be improved using deep learning techniques and real-time data from IoT devices. Mobile app support and multilingual features can also be added to enhance usability

IX. REFERENCES

[1] Singh, R., & Gupta, S. (2022). Hybrid Feature Selection and Random Forest for Thyroid Disease Prediction. Journal of Healthcare Engineering, 2022, 1–10.

[2] Ahmed, M., & Rahman, S. (2023). Early Detection of Thyroid Disorders Using Machine Learning Techniques. IEEE Access, 11, 34567–34578.

[3] Chen, Y., & Wang, L. (2021). Application of Random Forest Algorithm in Medical Diagnosis Systems. Journal of Biomedical Informatics, 118, 103789.

[4] Patel, K., & Shah, D. (2022). Comparative Analysis of Machine Learning Algorithms for Thyroid Disease Prediction. International Journal of Advanced Computer Science and Applications, 13(5), 210–217.

[5] Verma, S., & Mehta, R. (2021). Thyroid Disease Detection Using Machine Learning Approaches. Procedia Computer Science, 192, 3482–3490.

[6] Zhang, H., & Li, J. (2020). Feature Selection Techniques for Medical Data Classification. Expert Systems with Applications, 159, 113–121.

[7] Khan, M., & Ali, S. (2022). Random Forest Based Predictive Model for Healthcare Applications. Journal of Artificial Intelligence in Medicine, 126, 102231.