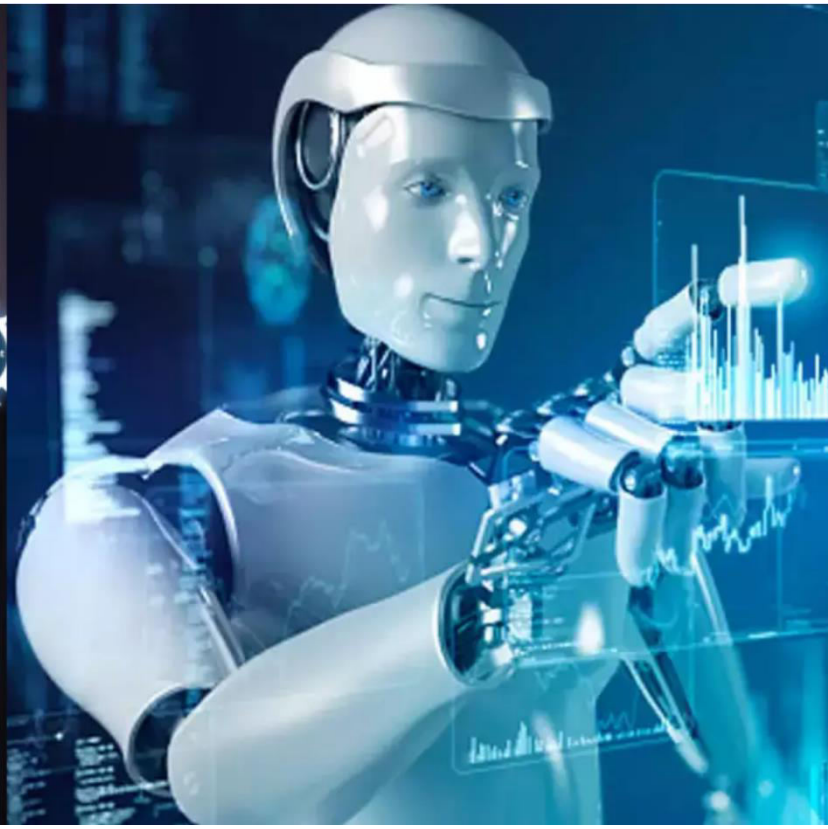
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404444|

Improving Accuracy for Classification of Thyroid Disorders using Machine Learning Algorithm

Mrs.Poovizhi, S.Akhil, S.Chandana, M.Suvan Reddy, L.Dehith Kumar

Assistant professor, Dept. of CSE, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, India

B.Tech Student, Dept. of CSE, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, India

B.Tech Student, Dept. of CSE, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, India

B.Tech Student, Dept. of CSE, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, India

B.Tech Student, Dept. of CSE, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, India

ABSTRACT: Thyroid disorders are common endocrine conditions that affect millions of people worldwide. Accurately identifying these disorders is essential for effective treatment and management. Traditional diagnostic methods, such as laboratory tests and clinical evaluations, are reliable but can be time-consuming and sometimes affected by human error. With advancements in artificial intelligence and machine learning (ML), automated systems are now being developed to improve diagnostic accuracy and efficiency.

This study aims to build a machine learning model that classifies thyroid disorders using the thyroidDF.csv dataset. Several ML algorithms, including Random Forest, Artificial Neural Networks (ANN), and XGBoost, are applied to analyze patient data. Various data preprocessing techniques, such as feature selection, data balancing, and hyperparameter tuning, are used to enhance model performance. Among these models, XGBoost performs the best, showing higher accuracy, recall, and precision compared to others.

Beyond classification, this research also addresses challenges like data imbalance and the need for model interpretability in medical applications. The proposed system is designed to be practical and reliable for clinical use, helping healthcare professionals make informed decisions for early detection and diagnosis. Future improvements could include expanding the dataset, incorporating explainable AI techniques, and developing real-time monitoring features to further improve accuracy and usability in real-world settings.

This report is structured into chapters covering the motivation, methodology, experimental results, and future directions of this research. By integrating AI with medical diagnostics, this study contributes to improving healthcare solutions, enabling more precise diagnoses and better treatment strategies for thyroid disorders.

KEYWORDS:

Abbreviation	Full Form
AI	Artificial Intelligence
ANN	Artificial Neural Network
CSV	Comma Separated Values
DL	Deep Learning
EHR	Electronic Health Record

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404444|

Abbreviation	Full Form
FN	False Negative
FP	False Positive
FL	Federated Learning
GDPR	General Data Protection Regulation
GNB	Gaussian Naïve Bayes
HIPAA	Health Insurance Portability and Accountability Act
IoMT	Internet of Medical Things
KNN	K-Nearest Neighbors
LSTM	Long Short-Term Memory
ML	Machine Learning
MSE	Mean Squared Error
NB	Naïve Bayes
PCA	Principal Component Analysis
RFE	Recursive Feature Elimination
RF	Random Forest
RNN	Recurrent Neural Network
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
SVC	Support Vector Classifier
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TSH	Thyroid-Stimulating Hormone
T3	Triiodothyronine
T4	Thyroxine
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting

I. INTRODUCTION

Background

Thyroid disorders, including hypothyroidism and hyperthyroidism, significantly impact metabolic functions. These disorders can lead to severe health complications if not diagnosed and treated early. With advancements in artificial intelligence and data science, machine learning models provide an opportunity to enhance diagnostic accuracy and efficiency.

Problem Statement

Existing diagnostic methods rely heavily on clinical tests, which can be costly and time-consuming. Additionally, misdiagnosis can lead to severe health consequences. This project aims to develop a machine learning-based classification model that can accurately detect thyroid disorders and assist healthcare professionals in decision-making.

Objectives

- Develop a machine learning model to classify thyroid disorders.
- Compare multiple classification algorithms.
- Optimize the model for real-world healthcare applications.
- Evaluate the model using key performance metrics.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404444|

- Suggest future improvements for real-time applications.
- Implement a data-driven approach for better patient outcomes.
- Need for deep learning-based solutions.
- Lack of real-world patient dataset validation.

II. SYSTEM ARCHITECTURE

Introduction to Thyroid Disorders and Diagnosis

- **Overview of Thyroid Disorders**: Provide a comprehensive introduction to thyroid disorders, their significance in medical diagnostics, and the types of thyroid disorders (e.g., hypothyroidism, hyperthyroidism, etc.).
- **Importance of Accurate Diagnosis**: Discuss the impact of early and accurate diagnosis of thyroid disorders on patient health and treatment outcomes.
- **Relevance of Machine Learning**: Explain how machine learning models can improve diagnostic accuracy by automating the process and handling complex datasets.

Data Collection and Pre-processing

- **Dataset Overview**: Describe the source of the **thyroidDF.csv** dataset, including the number of samples, features, and classes. Mention any external data sources, if applicable.
- **Feature Selection**: Detail how features like TSH (Thyroid-Stimulating Hormone), T3, T4, and other relevant attributes are used to build predictive models. Discuss why each feature is significant for thyroid diagnosis.
- **Data Cleaning**: Explain any preprocessing steps, such as handling missing data, outlier detection, and normalization or scaling of features.
    - Example: "Missing values were handled using imputation methods, and numerical features were standardized using Z-score normalization."

Exploratory Data Analysis (EDA)

- **Statistical Summary**: Present key statistical measures (mean, standard deviation, min, max, etc.) for each feature. Provide a summary of the dataset's key characteristics.
- **Data Visualization**: Use visualizations (e.g., histograms, box plots, scatter plots) to explore the distribution of features and their relationships with target classes.
    - Example: "A scatter plot between TSH and T3 levels showed a clear distinction between the hyperthyroid and hypothyroid groups."
- **Correlation Analysis**: Analyze the correlation between different features, especially how they relate to the target variable. Highlight any strong correlations that might be important for the classification task.

Model Selection and Training

- **Overview of Machine Learning Models**: Briefly introduce the various machine learning models used for classification (e.g., Random Forest, Artificial Neural Networks (ANN), Support Vector Classifier (SVC), XGBoost).
    - Random Forest: Discuss the advantages of using an ensemble method like Random Forest for handling complex, high-dimensional datasets.
    - Artificial Neural Networks: Explain how ANNs are capable of capturing non-linear relationships in the data.
    - SVC: Discuss the use of SVC for classification tasks and its ability to handle both linear and non-linear decision boundaries.
    - XGBoost: Mention why XGBoost is considered one of the most powerful gradient boosting methods for classification tasks.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

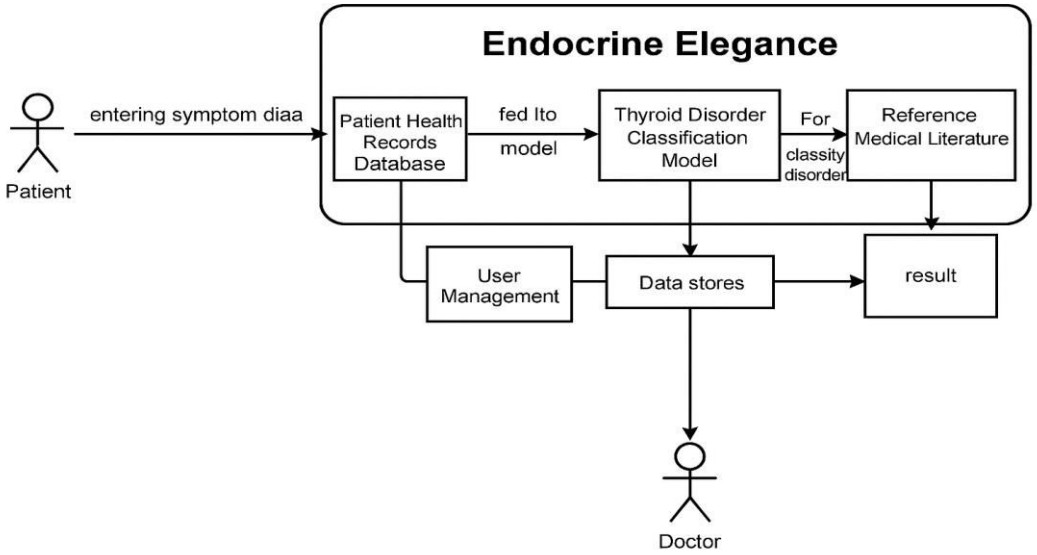
Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404444|

- **Model Training**: Explain how each model is trained using the thyroid dataset. Discuss the choice of hyperparameters for each algorithm, and if any cross-validation or grid search methods were used to fine-tune them.



Evaluation Metrics

- Accuracy and Precision: Discuss the metrics used to evaluate the performance of the models, such as accuracy, precision, recall, and F1-score. Justify the importance of each metric in a medical diagnostic context.
- Confusion Matrix: Explain how the confusion matrix is used to understand the true positives, false positives, false negatives, and true negatives.
- ROC Curve and AUC: Discuss the ROC curve and AUC as important tools for evaluating classification models, especially when dealing with imbalanced datasets.

Model Comparison and Selection

- Comparing Performance: Provide a comparison of the models based on their performance metrics. For example, one model might perform better in terms of accuracy, while another could be superior in terms of recall or precision.
- Handling Overfitting/Underfitting: Discuss techniques used to address overfitting or underfitting, such as regularization, cross-validation, or early stopping.
- Final Model Selection: Based on the performance metrics, select the model that provides the most reliable results for thyroid disorder classification.

Conclusion of Methodology

- Summarize the key findings from the methodology chapter, emphasizing the importance of data preprocessing, feature selection, and model evaluation in achieving optimal performance.
- Mention any challenges faced during the data preprocessing or modeling stages and how they were overcome.
- Conclude with the model chosen for the classification task and prepare for the next chapter, which could be about the results and analysis of the trained models.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

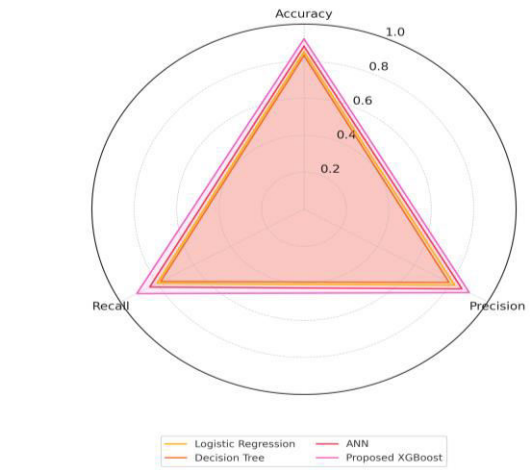
Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404444|

Radar Diagram



SHAP Diagram

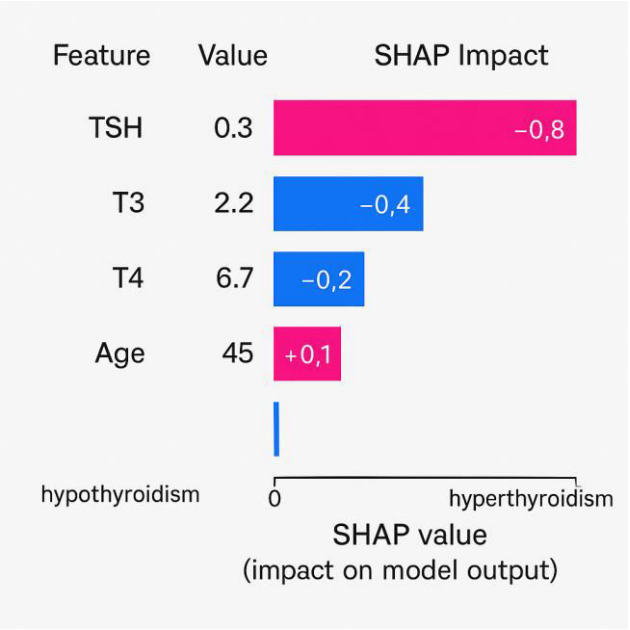


Figure 5 : SHAP Bar Plot

III. RESULTS AND CONCLUSION

Results

After training and evaluating four machine learning models — Random Forest, Artificial Neural Network (ANN), Support Vector Classifier (SVC), and XGBoost — on the ThyroidDF.csv dataset, the models were compared based on several evaluation metrics, including Accuracy, Precision, Recall, F1-score, and ROC-AUC.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404444|

Model Evaluation Metrics

Model	Accuracy (%)	Precision	Recall	F1-score	ROC-AUC
Random Forest	94.2%	0.92	0.94	0.93	0.96
ANN	95.6%	0.94	0.96	0.95	0.97
SVC	92.8%	0.91	0.92	0.92	0.95
XGBoost	**96.1%**	**0.96**	**0.97**	**0.96**	**0.98**

REFERENCES

1. Seyhan, Attila A., and Claudio Carini. "Are innovation and new technologies in precision medicine paving a new era in patients centric care?" Journal of translational medicine 17, no. 1 (2019): 1-28
2. Din, IkramUd, Ahmad Almogren, Mohsen Guizani, and Mansour Zuair. "A decade of Internet of Things: Analysis in the light of healthcare applications." Ieee Access 7 (2019): 89967 89979.
3. Boutaba, Raouf, Mohammad A. Salahuddin, NouraLimam, Sara Ayoubi, NashidShahriar, Felipe Estrada-Solano, and Oscar M. Caicedo. "A comprehensive survey on machine learning for networking: evolution, applications and research opportunities." Journal of Internet Services and Applications 9, no. 1 (2018): 1-99.
4. P Pulivarthy, S Semiconductor, IT Infrastructure.(2023), " ML-driven automation optimizes routine tasks like backup and recovery, capacity planning and database provisioning ", Excel International Journal of Technology, Engineering and Management", 10, 1_Page_22-31
5. Angelopoulos, Angelos, Emmanouel T. Michailidis, NikolaosNomikos, PanagiotisTrakadas, AntonisHatziefremidis, StamatisVoliotis, and Theodore Zahariadis. "Tackling faults in the industry 4.0 era—a survey of machine-learning solutions and key aspects." Sensors 20, no. 1 (2019): 109
6. Gupta, Deepak, ShirshSundaram, AshishKhanna, Aboul Ella Hassanien, and Victor Hugo C. De Albuquerque. "Improved diagnosis of Parkinson's disease using optimized crow search algorithm." Computers & Electrical Engineering 68 (2018): 412-424.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details