A Project report on

Thyroid Disease Classification Using Machine Learning

A Dissertation submitted to JNTUH, Hyderabad in partial fulfillment of the academic requirements for the award of the degree.

Bachelor of Technology

in

Artificial Intelligence and Machine Learning

Submitted by

D. KULDEEP
(21H51A7301)

K. VARUN KRISHNA
(21H51A7304)

P. YASHWANTH
(21H51A7307)

Under the esteemed guidance of

Mr. ANIL KUMAR
(Assistant Professor)

Department of Artificial Intelligence and Machine Learning

CMR COLLEGE OF ENGINEERING& TECHNOLOGY
(UGC Autonomous)
*Approved by AICTE *Affiliated to JNTUH *NAAC Accredited with A$^+$ Grade

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD - 501401.

2024-2025

.

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD – 501401

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

CERTIFICATE

This is to certify that the Major Project Phase-1 report entitled **"Thyroid Disease Classification using Machine Learning"** being submitted by D. Kuldeep(21H51A7301),K.VarunKrishna(21H51A7304),P.Yashwanth(21H51A7307) in partial fulfillment for the award of **Bachelor of Technology in Artificial Intelligence and Machine Learning** is a record of bonafide work carried out his/her under my guidance and supervision.

The results embodies in this project report have not been submitted to any other University or Institute for the award of any Degree.

Mr. Anil Kumar
Assistant Professor
Dept.of AIML

Dr. S.Kirubakaran
Professor and HOD
Dept. of AIML

ACKNOWLEDGEMENT

With great pleasure we want to take this opportunity to express my heartfelt gratitude to all the people who helped in making this project work a grand success.

We are grateful to **Mr. Anil Kumar ,Assistant Proffesor** , Department of Artificial Intelligence and Machine Learning for his valuable technical suggestions and guidance during the execution of this project work.

We would like to thank **Dr. S.Kirubakaran,** Head of the Department of Artificial Intelligence and Machine Learning, CMR College of Engineering and Technology, who is the major driving forces to complete my project work successfully.

We would like to thank **Dr. P. Ravi Kumar**, Dean F&S, CMR College of Engineering and Technology, for his insight and expertise have been instrumental in shaping the direction and execution of this project work successfully.

We are very grateful to **Dr. Ghanta Devadasu**, Dean-Academics, CMR College of Engineering and Technology, for his constant support and motivation in carrying out the project work successfully.

We are highly indebted to **Dr. V A Narayana,** Principal, CMR College of Engineering and Technology, for giving permission to carry out this project in a successful and fruitful way.

We would like to thank the Teaching & Non- teaching staff of Department of Computer Science and Engineering for their co-operation

We express our sincere thanks to **Shri. Ch. Gopal Reddy**, Secretary& Correspondent, CMR Group of Institutions, and **Shri Ch Abhinav Reddy**, CEO, CMR Group of Institutions for their continuous care and support

Finally, We extend thanks to our parents who stood behind us at different stages of this Project. We sincerely acknowledge and thank all those who gave support directly and indirectly in completion of this project work.

D. Kuldeep	21H51A7301
K. Varun Krishna	21H51A7304
P. Yashwanth	21H51A7307

TABLE OF CONTENTS

List of Figures

List of Tables

ABSTRACT

Thyroid disease classification using machine learning (ML) has gained attention as a promising method to enhance diagnostic accuracy and efficiency in clinical settings. This study focuses on the development and evaluation of various ML models for classifying thyroid disorders, including hyperthyroidism, hypothyroidism, and normal thyroid function. A comprehensive dataset is used, incorporating patient demographics, medical history, and laboratory test results. The methodology includes extensive data pre-processing to address missing values, normalize features, and ensure the quality of the dataset for model training. Several ML algorithms are implemented, such as Support Vector Machines (SVM), Random Forest (RF), and Neural Networks (NN), and are rigorously trained and validated using cross-validation techniques to guarantee robust and unbiased performance evaluation. To assess the efficacy of the models, evaluation metrics including accuracy, precision, recall, and F1-score are employed. The results indicate that ML models can effectively classify thyroid disorders, achieving high performance in terms of predictive accuracy. Furthermore, the study highlights the potential of machine learning to significantly improve diagnostic capabilities, offering an efficient tool to aid clinicians in early detection and better management of thyroid diseases. This research underscores the growing role of AI-driven approaches in healthcare, with the potential to enhance clinical decision-making and patient outcomes.

CHAPTER 1
INTRODUCTION

CHAPTER 1
INTRODUCTION

1.1 Problem Statement

Thyroid disorders, including hyperthyroidism, hypothyroidism, and normal thyroid function, are prevalent conditions that can significantly impact health if left undiagnosed or improperly managed. Traditional diagnostic methods often rely on clinical symptoms and laboratory tests, which may not always provide timely or accurate diagnoses. Furthermore, the increasing complexity and volume of medical data can overwhelm healthcare professionals, leading to potential delays or inaccuracies in diagnosis. Machine learning (ML) techniques have emerged as a promising solution to improve diagnostic accuracy by leveraging large datasets and advanced algorithms for classification. However, the application of ML to thyroid disease classification has yet to be fully explored in terms of its potential to enhance early diagnosis and assist clinicians in decision-making. This study aims to bridge this gap by developing and evaluating various ML models to classify thyroid conditions effectively.

1.2 Research Objective

The primary objective of this research is to develop, evaluate, and compare machine learning models for the classification of thyroid diseases. The specific objectives are as follows:

1. To explore the use of machine learning algorithms, such as Support Vector Machines (SVM), Random Forest (RF), and Neural Networks (NN), in the classification of thyroid conditions, including hyperthyroidism, hypothyroidism, and normal thyroid function.
2. To assess the performance of these models based on key evaluation metrics, including accuracy, precision, recall, and F1-score.
3. To determine the impact of data preprocessing techniques, such as handling missing values and feature normalization, on the performance of the models.

1.3 Project Scope and Limitations

Scope:

- **Dataset**: This study utilizes a comprehensive dataset that includes patient demographics, medical history, and laboratory test results, which are essential for accurate thyroid disease classification.
- **Machine Learning Models**: The research focuses on evaluating multiple machine learning algorithms, including Support Vector Machines (SVM), Random Forest (RF), and Neural Networks (NN), ensuring a wide exploration of potential models for thyroid disease classification.
- **Evaluation Metrics**: The models are assessed using common performance metrics like accuracy, precision, recall, and F1-score to ensure robust comparison and assessment of their efficacy.

Limitations:

- **Data Quality and Completeness**: The effectiveness of the machine learning models is highly dependent on the quality of the dataset. Missing data, errors in lab results, or inconsistencies in patient information can affect the accuracy of the models.
- **Model Generalization**: While cross-validation techniques are used to validate the models, their generalizability to other datasets or clinical settings remains a challenge, as the models are trained on a specific dataset.
- **Clinical Implementation**: The transition of these ML models into routine clinical practice may face challenges related to integration with existing healthcare systems, data privacy concerns, and the need for clinician training.
- **Focus on Specific Disorders**: This study focuses on the classification of hyperthyroidism, hypothyroidism, and normal thyroid function. It does not cover other thyroid-related conditions, such as thyroid cancer or autoimmune diseases, which could be areas for future research.

CHAPTER 2
BACKGROUND WORK

CHAPTER 2

BACKGROUND WORK

2.1. Existing Method 1: Thyroid Detection using Machine Learning

2.1.1. Introduction

The paper discusses the application of various machine learning algorithms, including Logistic Regression, Support Vector Machines (SVM), Decision Tree, Artificial Neural Networks (ANN), and k-Nearest Neighbors (KNN), for predicting thyroid diseases. These algorithms analyze thyroid blood test data to classify patients into categories such as normal, hyperthyroidism, or hypothyroidism. Data preprocessing is a critical step to ensure the training and testing datasets are clean and suitable for analysis.

2.1.2. Implementation of Existing Method 1

The machine learning models were trained using a dataset from the UCI repository. Steps included:

1. **Data Preprocessing**: Removing null or unnecessary data, splitting the dataset into training and test subsets.
2. **Feature Extraction**: Identifying key attributes (e.g., T3 resin uptake, total T4, TSH levels).
3. **Model Training and Evaluation**:
   - KNN: 93.84% accuracy.
   - SVM: 95.38% accuracy.
   - ANN: 75.38% accuracy.
   - Decision Tree: 92.3% accuracy.
   - Logistic Regression: 96.92% accuracy.
4. **Web Application Development**: Python Flask was used for backend integration with a web interface that accepts blood test data from users and predicts disease type.

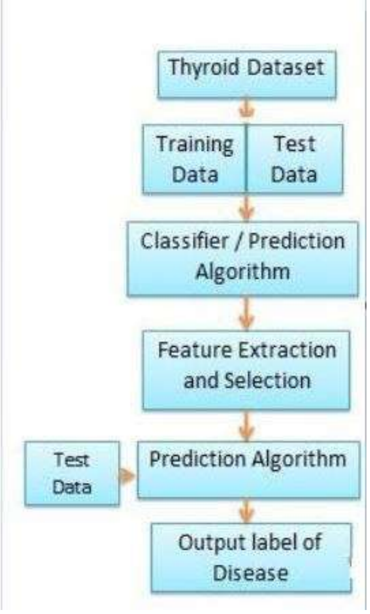

Fig 2.1: work flowchart

2.1.3. Merits, Demerits, and Challenges

- **Merits**:
  - Machine learning provides a precise and efficient method for disease detection.
  - Algorithms like Logistic Regression and SVM achieve high accuracy (e.g., Logistic Regression: 96.92%).
  - Can assist doctors by reducing the burden of manual diagnosis.
- **Demerits**:
  - The accuracy of prediction varies across algorithms (e.g., ANN shows lower accuracy of 75.38%).
  - Requires a clean and well-processed dataset; raw data with null values or noise can lead to poor results.

- **Challenges**:
  - Handling imbalanced data, as most datasets have a majority of normal cases and fewer hyperthyroidism or hypothyroidism cases.
  - Ensuring the generalizability of the model to different patient populations.
  - Development of a web interface that seamlessly integrates the machine learning model.

Classifier	Accuracy
KNN	93.84
SVM	95.38
ANN	75.38
Decision Tree	92.1
Logistic Regression	96.92

Table 2.1 Accuracy of algorithms.

2.2. Existing Method 2: Thyroid Detection Classification using Machine Learning

2.2.1. Introduction

The paper focuses on utilizing machine learning algorithms such as Decision Tree (J48), Decision Stump, and others for the classification of thyroid diseases. These models aim to distinguish between categories like negative, compensated hypothyroid, primary hypothyroid, and secondary hypothyroid using blood test data. Data preprocessing, including dimensionality reduction, plays a vital role in improving model performance and accuracy.

2.2.2. Implementation of Existing Method 2

The implementation involves the following steps:

1. **Data Preprocessing**:
   - Dataset: UCI Thyroid Dataset with 3772 instances and 30 attributes.
   - Dimensionality reduction to select 12 key attributes (e.g., TSH, T3, TT4).
2. **Model Training and Evaluation**:
   - **Decision Stump**:
     - Accuracy: 95.38%.
     - Simplicity: One-level decision tree.
   - **J48 Algorithm**:
     - Accuracy: 99.57%.
     - Recursive splitting of data using information gain.
3. **Performance Metrics**:
   - Confusion Matrix: Measures true positives, true negatives, false positives, and false negatives.
   - Accuracy and Error Rate: Evaluated for each classifier.
4. **Tools Used**:
   - WEKA for model development, evaluation, and visualization.

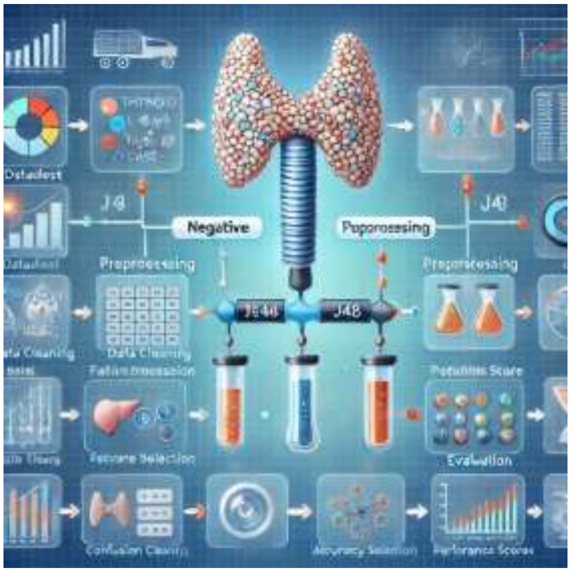

Fig:2.2 Process Followed

2.2.3. Merits, Demerits, and Challenges

- **Merits**:
  - High accuracy achieved using J48 (99.57%).
  - Preprocessing ensures reliable feature selection and dimensionality reduction.
  - Open-source tools like WEKA simplify implementation.
- **Demerits**:
  - Decision Stump has lower accuracy (95.38%) and limited predictive power.
  - Dataset imbalance leads to potential biases in model predictions.
- **Challenges**:
  - Developing models that generalize well across diverse datasets.
  - Reducing error rates while maintaining simplicity in decision-making.
  - Integrating models into real-time healthcare systems for practical use.

Classifier	Accuracy
Decision Stump	95.38
J48	99.57

Table 2.2. Accuracy of algorithms.

2.3. Existing Method 3: Optimized Machine Learning for Thyroid Detection

2.3.1. Introduction

This study leverages machine learning and deep learning approaches for multi-class thyroid disease detection, addressing issues like class imbalance and model optimization. The dataset consists of 9172 samples with ten thyroid-related classes. A Differential Evolution (DE) optimization algorithm is employed to fine-tune hyperparameters, enhancing performance. To resolve the class imbalance, Conditional Generative Adversarial Networks (CTGAN) are used for data augmentation.

2.3.2. Implementation of Existing Method 1

The implementation involves the following steps:

1. **Data Preparation**:
   - Dataset from Kaggle includes 31 features and 10 classes (e.g., hyperthyroid, hypothyroid, etc.).
   - CTGAN is applied to balance the dataset.
2. **Preprocessing**:
   - Data is encoded using Label Encoder to convert it into numeric form.
   - Dataset split into 80% training and 20% testing.
3. **Model Training and Optimization**:
   - Models trained include Random Forest (RF), Gradient Boosting Machine (GBM), AdaBoost, Support Vector Machine (SVM), and Logistic Regression (LR).
   - Differential Evolution (DE) is used to optimize hyperparameters like max_depth, n_estimators, learning rate, etc.

4. **Results**:
   - AdaBoost (Optimized): 99.8% accuracy.
   - GBM (Optimized): 99.6% accuracy.
   - RF (Optimized): 99.5% accuracy.

Performance Summary:

- **Metrics**: Models were evaluated using precision, recall, F1 score, and accuracy.
- **Confusion Matrix**: Shows significant improvements after optimization.

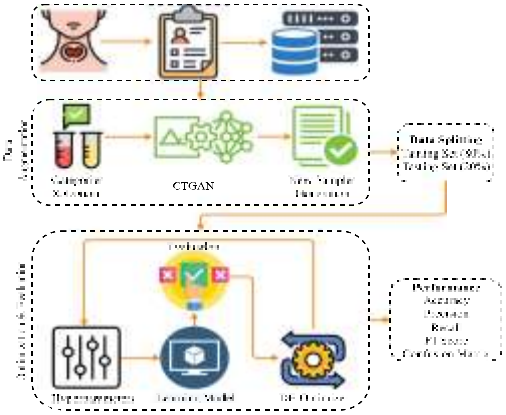

Fig:2.3.Work Flow

2.3.3. Merits, Demerits, and Challenges

- **Merits**:
  - Effective use of DE optimization for hyperparameter tuning.
  - CTGAN resolves class imbalance, reducing overfitting and improving predictions.
  - Models demonstrate high accuracy (e.g., AdaBoost: 99.8%).
- **Demerits**:
  - Computational cost of CTGAN and DE optimization.
  - Logistic Regression underperforms (64.3% accuracy).

Challenges:

- Balancing imbalanced datasets effectively while preserving feature distribution.
- Generalizing multi-class predictions across diverse thyroid disease cases.

Classifier	Accuracy
AdaBoost	99.8
GBM	99.6
RandomForest	99.5
SVM	96.6
LR	64.3

Table 2.3. Accuracy of algorithms

CHAPTER 3

RESULTS AND DISCUSSION

CHAPTER 3

RESULTS AND DISCUSSION

3.1. Comparison of Existing Solutions

The comparison of existing machine learning approaches for thyroid disease classification highlights the strengths and limitations of various algorithms:

[1] **High-Performance Models**:

- **AdaBoost** with Differential Evolution optimization achieved the highest accuracy (99.8%), proving highly effective for multi-class classification tasks.
- **Decision Tree (J48)** was close behind with 99.57%, demonstrating simplicity and reliability for structured datasets.
- **Logistic Regression** showed good performance (96.92%) for binary and smaller datasets, with ease of implementation.

[2] **Deep Learning Models**:

- CNN and LSTM models underperformed compared to traditional machine learning models for structured datasets, achieving lower accuracies. However, these methods are well-suited for large-scale datasets and image-based detection tasks.

[3] **Challenges in Existing Approaches**:

- **Class Imbalance**: Many existing solutions struggled with imbalanced datasets, affecting their ability to predict minority class outcomes accurately.
- **Overfitting**: Some models overfitted the majority class, leading to reduced generalizability.

3.2. Data Collection and Performance Metrics

1. **Data Collection**:
   - **Datasets**: Most studies used publicly available datasets from UCI and Kaggle repositories.
   - **Sample Characteristics**:
     - Kaggle dataset: 9172 instances, 10 target classes.
     - UCI dataset: 3772 instances with a focus on binary and multi-class thyroid disease classification.
   - **Preprocessing**: Data cleaning, feature selection, and dimensionality reduction were critical steps to improve model performance.
2. **Performance Metrics**:
   - **Accuracy**:
     - AdaBoost: 99.8%.
     - J48 Decision Tree: 99.57%.
     - Logistic Regression: 96.92%.
   - **Precision, Recall, and F1 Scores**: These metrics provided a balanced evaluation of model performance, particularly for imbalanced datasets.
   - **Confusion Matrix**: Used extensively to evaluate classification errors and assess true/false positives and negatives.
3. **Techniques for Improvement**:
   - **Data Augmentation**: CTGAN addressed class imbalance by generating synthetic samples for minority classes.
   - **Hyperparameter Tuning**: Differential Evolution significantly enhanced model performance by optimizing parameters like learning rate and tree depth.

CHAPTER 4

CONCLUSION

CHAPTER 4

CONCLUSION

The comparison of machine learning solutions for thyroid disease detection highlights the importance of selecting suitable models, addressing data imbalances, and optimizing hyperparameters to achieve high accuracy.

Key findings include:

1. **High Accuracy Models**:
   - Ensemble methods like AdaBoost (99.8%) and Decision Tree (99.57%) outperformed others.
   - Logistic Regression achieved 96.92%, suitable for simpler datasets.
2. **Deep Learning vs. Machine Learning**:
   - Machine learning models generally outperformed deep learning methods like CNN and LSTM for smaller datasets, though deep learning is advantageous for image-based tasks.
3. **Challenges Addressed**:
   - Techniques like CTGAN resolved class imbalance, while Differential Evolution (DE) reduced overfitting.

In conclusion, integrating advanced machine learning models with optimization techniques provides an effective approach to thyroid disease detection. Future work can expand detection capabilities by combining blood test data with image processing, addressing complex cases such as thyroid nodules and cancers. This framework offers significant potential for improving healthcare diagnostics.

CHAPTER 5

REFERENCES

CHAPTER 5
REFERENCES

[1] **Gupta, P., Rustam, F., Kanwal, K., et al.**
Detecting Thyroid Disease Using Optimized Machine Learning Model Based on Differential Evolution.
International Journal of Computational Intelligence Systems, 2024. DOI: 10.1007/s44196-023-00388-2.

[2] **Ram Kumar, R. P., Sri Lakshmi, M., Ashwak, B. S., et al.**
Thyroid Disease Classification Using Machine Learning Algorithms.
E3S Web of Conferences, 2023. DOI: 10.1051/e3sconf/202339101141.

[3] **Chandan R., Vasan, C., Chethan, M. S., et al.**
Thyroid Detection Using Machine Learning.
International Journal of Engineering Applied Sciences and Technology, 2021. Vol. 5, Issue 9, Pages 173-177. Available at: IJEAST.

[4] **Tyagi, A., Mehra, R.**
Interactive Thyroid Disease Prediction System Using Machine Learning Techniques.
IEEE International Conference on Parallel, Distributed, and Grid Computing (PDGC), 2018.

[5] **Godara, S.**
Prediction of Thyroid Disease Using Machine Learning Techniques.
International Journal of Electrical Engineering, 2018.

[6]International Journal of Computational Intelligence Systems
Thyroid Disease Prediction with Class Balancing and Model Optimization.
Comprehensive methodology for addressing data imbalance and enhancing prediction accuracy using AdaBoost and CTGAN

[7] **Ozyılmaz, L., Yıldırım, T.**
Diagnosis of Thyroid Disease Using Artificial Neural Network Methods.
ICONIP, Proceedings of the 9th International Conference on Neural Information Processing, 2002.

[8]E3S Web of Conferences
Thyroid Disease Detection Using Machine Learning.
Analysis based on datasets from UCI repository, highlighting the use of J48 and Decision Stump models for thyroid detection.