THYROID DETECTION USING LOGISTIC REGRESSION

A DESIGN PROJECT REPORT

Submitted by

AGNES MARY LAVANYA A (811721243004)

BRINDHA G (811721243012)

SAHANA SRI D (811721243046)

VINODHA R (811721243062)

in partial fulfilment for the award of the

degree of

BACHELOR OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENGE AND DATA SCIENCE

K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY

(An Autonomous Institution, affiliated to Anna University Chennai and Approved by AICTE,New Delhi)

SAMAYAPURAM–621112

DECEMBER-2023

K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY
(AUTONOMOUS)
SAMAYAPURAM–621112

BONAFIDE CERTIFICATE

Certified that this design project report titled **"THYROID DETECTION USING LOGISTIC REGRESSION"** is the bonafide work of **AGNES MARY LAVANYA A (811721243004), BRINDHA G (811721243012), SAHANA SRI D (811721243046), VINODHA R (811721243062)** who carried out the project under my supervision. Certified further, that to the best of my knowledge the work reported here in does not form part of any other project report or dissertation based on which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr. T. AVUDAIAPPAN M.E., Ph.D.,

HEAD OF THE DEPARTMENT,
Department of AI,
K. Ramakrishnan College of Technology
(Autonomous),
Samayapuram – 621 112.

SIGNATURE

Mrs. D. DEENA ROSE., M.E., (Ph.D.),

ASSISTANT PROFESSOR,
Department of AI,
K. Ramakrishnan College of Technology
(Autonomous),
Samayapuram – 621 112.

Submitted for the viva-voce examination held on ……

INTERNAL EXAMINER **EXTERNAL EXAMINER**

DECLARATION

We jointly declare that the project report on **"THYROID DETECTION USING LOGISTIC REGRESSION"** is the result of original work done by us and best of our knowledge, similar work has not been submitted to **"ANNA UNIVERSITY CHENNAI"** for the requirement of Degree of **BACHELOR OF TECHNOLOGY**. This design project report is submitted on the partial fulfilment of the requirement of the award of Degree of **BACHELOR OF TECHNOLOGY**.

SIGNATURE

AGNES MARY LAVANYA A

BRINDHA G

SAHANA SRI D

VINODHA R

PLACE : SAMAYAPURAM

DATE :

ACKNOWLEDGEMENT

It is with great pride that we express our gratitude and in - debt to our institution **"K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY (AUTONOMOUS)"**, for providing us with the opportunity to do this project.

We are glad to credit honourable chairman **Dr. K. RAMAKRISHNAN**, **B.E.,** for having provided for the facilities during the course of our study in college.

We would like to express our sincere thanks to our beloved Executive Director **Dr. S. KUPPUSAMY, MBA., Ph.D.,** for forwarding to our project and offering adequate duration in completing our project.

We would like to thank our principal **Dr. N. VASUDEVAN, M.E., Ph.D.,** who gave opportunity to frame the project the full satisfaction.

We whole heartily thanks to **Dr. T. AVUDAIAPPAN**, **M.E., Ph.D.,** HEAD OF THE DEPARTMENT, **ARTIFICAL INTELLIGENCE** for providing his encourage pursuing this project.

We express our deep and sincere gratitude to my project guide **Mrs. D. DEENA ROSE., M.E., (Ph.D.),** ASSISTANT PROFESSOR, **ARTIFICIAL INTELLIGENCE** for her incalculable suggestions, creativity, assistance and patience which motivated me to carry out the project successfully.

We render our sincere thanks to my project coordinator **Mr. R. ROSHAN JOSHUA M.E.,** other faculties and non-teaching staff members for providing valuable information during the course. We wish to express our special thanks to the officials & Lab Technicians of our departments who rendered their help during the period of the workprogress.

ABSTRACT

Thyroid gland is one of the body's most important glands because it regulates the metabolism of the human body. It controls how the body works by releasing specific hormones into the blood. The two different hormone disorders are hypothyroidism and hyperthyroidism. When these disorders occur, the thyroid gland releases a particular hormone into the blood that regulates the metabolism of the body. Iodine deficiency, autoimmune conditions, and inflammation can contribute to thyroid issues. The disease is diagnosed using a blood test, but there is frequently some noise and disturbance. Techniques for cleaning data can be used to make it simple enough to perform analytics that show the patient's risk of developing thyroid disease. Our proposed system deals with the logistic regression for analysis and classification models used in thyroid disease based on the information gathered from the dataset taken from the UCI machine learning repository. Machine learning plays a crucial role in the detection of thyroid disease. Our proposed system suggests logistic regression in machine-learning for thyroid detection and diagnosis for thyroid prevention.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

KNN	K-Nearest Neighbor
ML	Machine Learning
SVM	Support Vector Machine
T3	Tri-iodothyronine
T4	Thyroxine
TSH	Thyroid-stimulating hormone
TT4	Total Thyroxine
T4U	Thyroxine Utilization Rate

CHAPTER 1

INTRODUCTION

The diagnosis of thyroid disease fully depends on hormones. Generally doctors use medical history in diagnosis but it is not sufficient because without physical exam and medical hormonal test does not diagnose clearly. Thyroid function inter relate with every function in human body. Human body function not properly work then some symptoms overcome in human body as like fatigue, weight gain, mood issue, irregular period, muscle pain cold hand, dry and cracking skin neck etc., These cells absorb Iodine and amino acid tyrosine for creation T3 and t4, t3 and T4 control metabolism of the body. T3, T4 control and manage oxygen and calories and create it into energy. It is very important in human that T3 and T4 levels must be always in balancing order. Here logistic regression algorithm is employed for predicting thyroid disease besides using any other machine learning methods or algorithms. Because logistic regression provides better accuracy and efficiency for our model. This proposed system tries to find the combination of T3, T4 and TSH for good health.

1.1 BACKGROUND

Our proposed system investigates logistic regression's role in thyroid detection, focusing on early diagnosis. It reviews thyroid disorders, emphasizes logistic regression's significance in medical diagnostics, and details data collection and preprocessing steps. Methodologically, it covers feature selection, model training, and comprehensive evaluation metrics like accuracy, precision, recall, and F1-score. Results showcase model performance on training and validation sets, supplemented by critical analyses such as confusion matrices or ROC curves. The discussion interprets coefficients, assesses clinical implications, compares methodologies, and proposes enhancements. Conclusively, it emphasizes logistic regression's potential in early thyroid detection and highlights areas for further research in refining diagnostic models. With a global impact, the endeavor aims for enhanced accuracy in identifying thyroid abnormalities like hyperthyroidism. Through logistic regression and extensive medical datasets, the project refines classification, automates diagnostics, and accelerates healthcare efficiency. Initial strides demonstrate promising advancements in accuracy and classification precision, marking a significant shift in early detection for thyroid disorders.

1.2 PROBLEM STATEMENT:

The current diagnostic methods for thyroid disorders lack consistent accuracy and efficiency, demanding a more reliable approach. This proposed system seeks to develop a logistic regression model for thyroid abnormality detection based on medical data. The challenge lies in accurately differentiating between normal thyroid conditions and various abnormalities, including hyperthyroidism and hypothyroidism, using logistic regression techniques. The objective is to create a dependable algorithm that enhances diagnostic precision, reduces false positives, and contributes to efficient and accurate thyroid disorder identification, ultimately aiding healthcare professionals in timely interventions and improved patient care.

1.3 AIMS AND OBJECTIVES:

1.3.1 AIM:

- To significantly improve accuracy.
- To Construct a logistic regression algorithm to accurately classify thyroid conditions into distinct categories such as normal, hyperthyroidism and hypothyroidism.
- To Identify and select the most relevant features from medical data that contribute significantly to the differentiation of thyroid abnormalities.
- To Compare the efficiency of logistic regression against other machine learning approaches to showcase its superiority in thyroid disorder detection.
- To minimize false positives and false negatives, thereby improving overall diagnostic accuracy and reducing the chances of misclassification.
- To evaluate the model's performance in distinguishing between healthy individuals and those with thyroid abnormalities, ultimately contributing to the advancement of early diagnostic methods in the field of thyroid health".

1.3.2 OBJECTIVES:

- Train a logistic regression model to accurately classify thyroid conditions, such as normal, nodules, hyperthyroidism, and hypothyroidism, based on features extracted from medical imaging data.
- Ensure the developed model aligns with clinical diagnoses and is interpretable for healthcare professionals, aiding in decision-making for patient care and treatment strategies.
- Create a model that is computationally efficient and scalable for integration into medical practice, allowing for quick and reliable assessment of thyroid conditions.
- Validate the model's performance across diverse datasets to ensure its robustness and generalizability to different patient demographics and imaging variations.
- To increase overall accuracy by using logistic regression.

CHAPTER 2

LITERATURE SURVEY

2.1 TITLE: PREDICTION OF HYPOTHYROIDISM AND HYPERTHYROIDISM USING ML ALGORITHMS.

AUTHOR: Anika Shama, Md. Bipul Hossain, Apurba Adhikary, K. M. Aslam Uddin, Md. Amzad Hossain.

YEAR OF PUBLICATION: 2022

ALGORITHM USED: Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, Naïve Bayes Classifier, Logistic Regression, K-Nearest Neighbor and Support Vector Machine Algorithm.

ABSTRACT: In this study, we used a variety of machine learning algorithm to predict hypothyroidism and hyperthyroidism. Moreover, we identified the most significant features, which can be used to detect thyroid diseases more precisely. After completing the preprocessing and feature selection steps, we applied our modified and original data to several classification models to predict thyroidism. Finally, we found Random Forest is giving the maximum score in all sectors like accuracy, precision, recall, F1 score in our dataset and Naive Bayes is performing very poorly. In terms of accuracy and other performance evaluation criteria, this study could advocate the use of effective classifiers and features backed by machine learning algorithms for the detection and diagnosis of thyroid disease.

MERITS: ML algorithms can process vast amounts of data quickly, aiding healthcare professionals in making faster and more informed decisions.

DEMERITS: Models heavily rely on the quality and quantity of data. Incomplete or biased datasets can lead to inaccurate predictions or reinforce biases.

2.2 TITLE: INCREASING THE PREDICTION ACCURACY FOR THYROID DISEASE: A STEP TOWARDS BETTER HEALTH FOR SOCIETY.

AUTHOR: Ritesh Jha, Vandana Bhattacharjee, Abhijit Mustafi.

YEAR OF PUBLICATION: 2021

ALGORITHM USED: Principal Component Analysis, Decision Tree Classifier.

ABSTRACT: Thyroid diagnosis, particularly for an inexperienced clinician, is a difficult proposal. Many researchers have established various methods for the diagnosis of the disease and several models for disease prediction have been developed. As with several other domains, machine learning approaches to modelling health care problems is gaining popularity. This study aims at providing solutions towards such a thyroid disease prediction. Dimension reduction techniques are applied, and reduced dimension data input to classifiers. Also, data augmentation is applied so as to be able to generate sufficient data for deep neural network model. Classifier prediction is compared to other similar researches. Real life dataset for thyroid disease has been used, and experiments conducted in distributed environment. Our proposed two stage approach gives a maximum accuracy of 99.95% which is very good as compared to existing techniques. We have shown that dimension reduction and data augmentation can be used very efficiently for achieving high accuracy of disease prediction.

MERITS: With large datasets, machine learning models can achieve high accuracy in predicting thyroid disorders by considering various parameters like TSH, T3, T4 levels, symptoms, demographics, etc.

DEMERITS: ML models can overfit the training data, meaning they might perform well on training data but poorly on new, unseen data.

2.3 TITLE: THYROID DISEASE CLASSIFICATION USING MACHINE LEARNING ALGORITHMS.

AUTHOR: Khalid Salman, Emrullah Sonuc.

YEAR OF PUBLICATION: 2021

ALGORITHM USED: Support Vector Machine, Random Forest, Decision Tree, Naïve bayes, Logistic regression, K-Nearest Neighbors, Multi-Layer Perceptron and Linear Discriminant Analysis.

ABSTRACT: With the vast amount of data and information difficult to deal with, especially in the health system, machine learning algorithms and data mining techniques have an important role in dealing with data. In our study, we used machine learning algorithms with thyroid disease. The goal of this study is to categorize thyroid disease into three categories: hyperthyroidism, hypothyroidism, and normal, so we worked on this study using data from Iraqi people, some of whom have an overactive thyroid gland and others who have hypothyroidism, so we used all of the algorithms. Support vector machines, random forest, decision tree, naïve bayes, logistic regression, k-nearest neighbors, multi-layer perceptron (MLP), linear discriminant analysis.

MERITS: Enhanced accuracy and feature learning capabilities in classification tasks.

DEMERITS: Increased computational complexity and training time due to the combination of powerful algorithms.

2.4 TITLE: INTERACTIVE THYROID DISEASE PREDICTION SYSTEM USING MACHINE LEARNING TECHNIQUE.

AUTHOR: Ankita Tyagi, Ritika Mehra, Aditya Saxena.

YEAR OF PUBLICATION: 2018

ALGORITHM USED: Support Vector Machine, K-NN, Decision Tree Algorithm.

ABSTRACT: Hyperthyroidism and hypothyroidism are one of the two common diseases of the thyroid that releases thyroid hormones in regulating the rate of body's metabolism. Data cleansing techniques were applied to make the data primitive enough for performing analytics to show the risk of patients obtaining thyroid. The machine learning plays a decisive role in the process of disease prediction and this paper handles the analysis and classification models that are being used in the thyroid disease based on the information gathered from the dataset taken from UCI machine learning repository. It is important to ensure a decent knowledge base that can be entrenched and used as a hybrid model in solving complex learning task, such as in medical diagnosis and prognostic tasks. In this paper, we also proposed different machine learning techniques and diagnosis for the prevention of thyroid. Machine Learning Algorithms, support vector machine (SVM), K-NN, Decision Trees were used to predict the estimated risk on a patient's chance of obtaining thyroid disease.

MERITS: Effective modeling of temporal dependencies in sequential data.

DEMERITS: Computational complexity and data requirements can be high.

2.5 TITLE: AN EXPLAINABLE ARTIFICIAL FRAMEWORK FOR THE PREDICTIVE ANALYSIS OF HYPO AND HYPER THYROIDISM USING MACHINE LEARNING ALGORITHMS.

AUTHOR: Anika Shama, Md. Bipul Hossain, Apurba Adhikary, K. M. Aslam Uddin, Md. Amzad Hossain, Avi Deb Raha, Saydul Akbar Murad, Md. Shirajum Munir.

YEAR OF PUBLICATION: 2023

ALGORITHM USED: Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, Naïve Bayes Classifier, Logistic Regression, K-Nearest Neighbor, and Support Vector Machine (SVM) algorithms.

ABSTRACT: In this paper, we use different machine learning algorithms to predict hypothyroidism and hyperthyroidism. Moreover, we identified the most significant features, which can be used to detect thyroid diseases more precisely. After completing the pre-processing and feature selection steps, we applied our modified and original data to several classification models to predict thyroidism. We found Random Forest (RF) is giving the maximum evaluation score in all sectors in our dataset, and Naive Bayes is performing very poorly. Finally, we did some explainable analysis of our best classifier to understand the internal black-box of our machine learning model and datasets. This study could further pave the way for the researcher as well as healthcare professionals to analyze thyroid disease in real time applications.

MERITS: Real-time, non-invasive monitoring of tissue conductivity changes for medical applications.

DEMERITS: Increased resource requirements in terms of data and computation for improved accuracy.

CHAPTER 3

SYSTEM SPECIFICATION

3.1 HARDWARE CONFIGURATION:

- **Computer** - minimum of 4GB RAM & dual-core processor
- **RAM** – minimum 16GB to 32GB.
- **Stable internet connection**
- **Storage**

3.2 SOFTWARE CONFIGURATION:

- **Python programming language**
- **Operating system** - Windows, Linux, or macOS.
- **Python libraries** such as –numpy, pandas.

3.3 SOFTWARE DESCRIPTION:

For a thyroid detection project employing logistic regression, a suite of software tools is essential to handle data, build and train the model, and evaluate its performance.

3.3.1 COMPONENTS:

To develop the Thyroid Detection System here is a software description outlining the key components:

1.Python: Python serves as the primary programming language due to its extensive libraries and frameworks for machine learning and data analysis. Employ Python for developing a thyroid detection system due to its extensive libraries, like pandas for data handling and scikit-learn for machine learning.

2.NumPy and Pandas: NumPy offers support for numerical operations and arrays, while Pandas provides data manipulation tools with data structures like Data Frames. NumPy and Pandas streamline thyroid data handling, enabling efficient numerical operations, array handling, and comprehensive data organization for detection systems.

3.Scikit-learn: Scikit-learn offers a wide range of tools for machine learning tasks, including logistic regression implementation and model evaluation metrics. scikit-learn is used to preprocess, select features, train models, evaluate, tune hyperparameters, and create streamlined pipelines.

4.Matplotlib and Seaborn: Matplotlib and Seaborn are visualization libraries in Python, facilitating data visualization and model performance analysis.

5.Jupyter Notebooks or IDEs (Spyder, VS Code): Jupyter Notebooks provide an interactive environment, while IDEs like Spyder or Visual Studio Code offer code editing and debugging features.

3.4 Developing Environment

To develop the Hand Gesture Recognition System, you would typically set up thefollowing environment:

Python: Python is the primary programming language used for developing the system. Ensure that Python is installed on your system.

Integrated Development Environment (IDE): Choose an IDE for Python development, such as PyCharm, Visual Studio Code, or Jupyter Notebook. These IDEs provide features like code editing, debugging, and project management, enhancing the development process.

Virtual Environment: A virtual environment is a self-contained directory isolating Python interpreter and libraries for a specific project

Activation of Virtual Environment: Activating a virtual environment configures the terminal to use the specific Python interpreter and libraries of that environment.

Installation of Required Libraries: Libraries like NumPy, pandas, and scikit-learn are installed to provide essential tools for data manipulation and machine learning

Data Collection, Preprocessing, Train-Test Split: Collect relevant datasets, preprocess data by cleaning and transforming, and split it into training and testing sets.

CHAPTER 4

SYSTEM ANALYSIS

4.1 EXISTING SYSTEM

- In this work, including the k-Nearest-Neighbor algorithm, decision trees, support vector machines, and artificial neural networks. Classification and prediction based on the data set obtained from the UCI Repository were carried out, and accuracy was obtained based on the output produced.
- SVM machine learning and logistic regression are used Based on Precision, Recall, F measure, ROC, and RMS error, a comparison between these two algorithms was made. In the end, the best classifier was logistic regression.
- The classification accuracy, sensitivity, and specificity of the radiomics-based method are 66.81%, 51.19%, and 75.77%, respectively, while the evaluation indices for the deep learning-based method trained on the test samples are 74.69%, 63.10%, and 80.20%, respectively. The most effective methods ended up being deep learning.

ALGORITHM USED:

1. SUPPORT VECTOR MACHINES (SVM): While Support Vector Machines (SVM) are powerful tools for thyroid detection in machine learning, they come with certain drawbacks. One limitation is the sensitivity to the choice of hyperparameters, such as the regularization parameter (C) and the kernel type. Tuning these parameters can be computationally expensive and might require extensive experimentation to achieve optimal performance. Additionally, SVMs may not perform well on large datasets, as the time complexity increases with the square of the number of samples. Another drawback is that SVMs are binary classifiers, and extension to multiclass problems involves strategies like one-vs-one or one-vs-all, which can be less intuitive. Moreover, SVMs are less effective when dealing with noisy data or datasets with overlapping classes, as they aim to find a clear boundary between classes. Despite these drawbacks, SVMs remain a valuable tool for many classification tasks, including thyroid detection, especially when appropriate preprocessing and parameter tuning are applied.

2. DECISION TREE CLASSIFIER: Implementing a thyroid detection project using decision trees offers interpretability, simplicity in visualization, and ease of understanding for non-technical stakeholders. However, decision trees are prone to overfitting, especially when the tree depth is not properly regulated or when dealing with noisy or high-dimensional data. They might create overly complex trees that fail to generalize well to new, unseen data. Additionally, decision trees can be sensitive to small variations in the data, leading to different trees being generated with minor changes in the training set. Furthermore, they may not perform optimally compared to more sophisticated algorithms in certain scenarios, especially when dealing with intricate relationships among features or when handling imbalanced datasets, where they might favor the majority class. Therefore, while decision trees offer simplicity and interpretability, their performance might lag behind more complex models in terms of predictive accuracy and generalization.

3. K-NEAREST NEIGHBOR: The implementation of thyroid detection using the k-Nearest Neighbors (KNN) algorithm involves several key steps. Initially, a dataset containing relevant features such as medical test results and patient demographics is collected and preprocessed. The preprocessing step includes handling missing values and normalizing numerical features. The dataset is then split into training and testing sets. During the training phase, the KNN algorithm memorizes the feature vectors of the training instances. In the testing phase, for each instance in the testing set, the algorithm identifies the k-nearest neighbors in the training set based on a distance metric (e.g., Euclidean distance). The majority class among these neighbors determines the class of the test instance. The algorithm's performance is evaluated using metrics like accuracy and F1 score. KNN is particularly suitable for thyroid detection tasks when the underlying data exhibits local patterns, and the choice of an appropriate distance metric is crucial for its success.

1. DRAWBACKS:

- SVMs are sensitive to parameter choices and may require extensive tuning, making them computationally intensive. Their scalability is a concern for large datasets.
- KNN, although intuitive, becomes computationally expensive with high-dimensional or extensive data, and its predictions can be sensitive to noise and outliers. The memory usage is also a consideration.
- Decision Trees, while interpretable, are prone to overfitting, especially with deep trees, and may struggle with capturing intricate non-linear relationships. Additionally, they exhibit instability with minor data variations.

4.2 PROPOSED SYSTEM:

In case of using Decision tree algorithm, Support vector machine and K-Nearest Neighbor algorithm we may need to face with the complexity and overfitting since these algorithms are prone to overfitting. These can be computationally expensive, especially when dealing with a large number of trees and features. To overcome these issues we are giving a turn to logistic regression since it provides easily interpretable results and makes it easy to understand the importance of individual features in the diagnosis. Logistic regression is a simple and linear model, which is computationally effective and efficient.

ALGORITHM USED:

LOGISTIC REGRESSION ALGORITHM:

Logistic regression is a statistical method used in thyroid detection, predicting the probability of thyroid disorder based on input features. It models the relationship between the dependent binary outcome (presence or absence of thyroid disorder) and independent variables (clinical data or test results). By fitting a logistic curve to the data, the algorithm estimates the likelihood of thyroid dysfunction, providing a valuable tool for medical diagnostics and decision-making.

ADVANTAGES:

- Precise Probabilities
- Interpretability
- Robust Predictions
- Diagnostic Insights
- Efficiency

CHAPTER 5

ARCHITECTURAL DESIGN

5.1 SYSTEM DESIGN:

A system architecture is the conceptual model that defines the structure, behavior, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system.

A thyroid detection project employing logistic regression utilizes statistical modeling in Python with Scikit-learn. It involves data preprocessing, feature selection and model evaluation for binary classification of thyroid conditions.

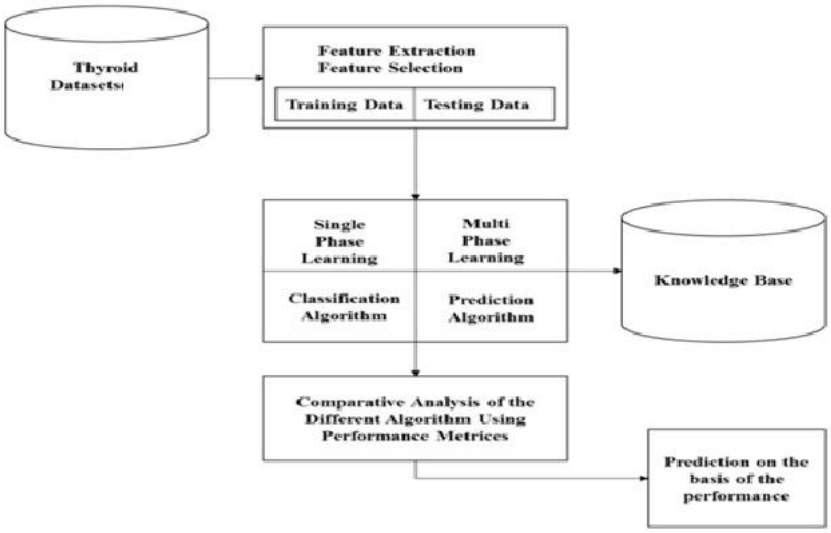

FIG 5.1: ARCHITECTURE DIAGRAM

5.2 DATA FLOW DIAGRAM:

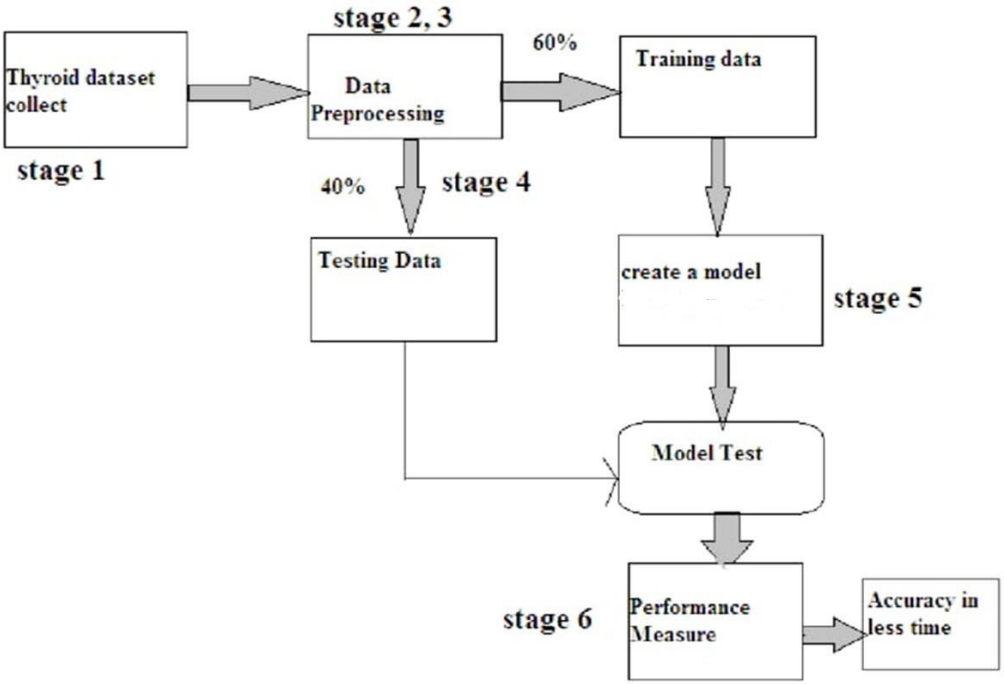

FIG 5.2 : DATA FLOW DIAGRAM

5.3 USE CASE DIAGRAM:

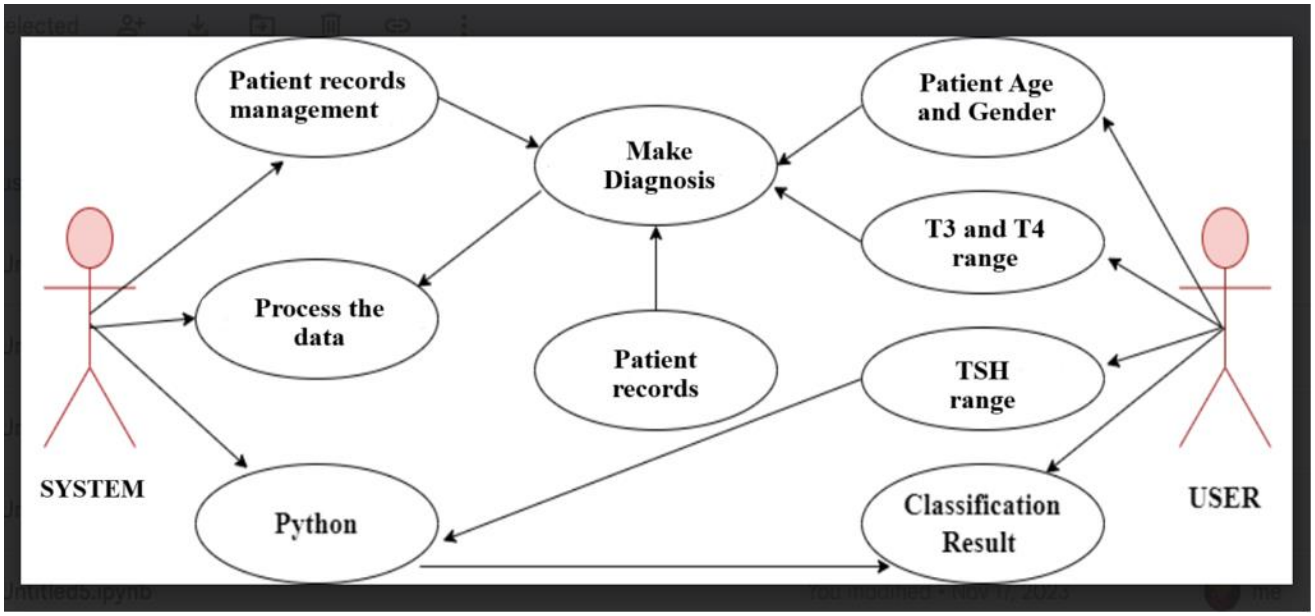

FIG 5.3 : USE CASE DIAGRAM

5.4 ACTIVITY DIAGRAM:

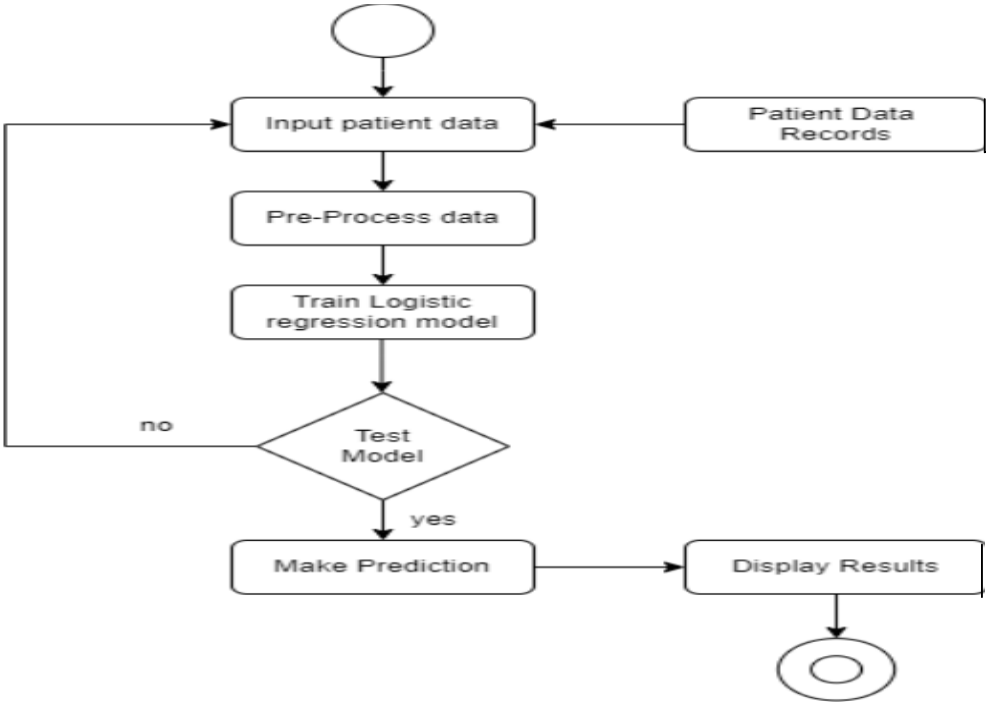

FIG 5.4 : ACTIVITY DIAGRAM

CHAPTER 6
MODULE DESCRIPTION

6.1 MODULES:

6.1.1 Data Import Module:

It plays a crucial role in the thyroid detection system, seamlessly integrating text-based data from CSV files. Utilizing Python libraries like Pandas, it ensures efficient handling and compatibility, facilitating flexible dataset updates for ongoing research.

Here are the key tasks involved in the Data Import module:

1. **Purpose:** Facilitates the incorporation of the thyroid dataset into the system.
2. **Functionality:** Reads data from a CSV file containing text-based information on thyroid conditions.
3. **Libraries Used:** Utilizes Python libraries like Pandas to efficiently import and manage tabular data.

6.1.2 Data Preprocessing Module:

The Data Preprocessing refines the imported data for logistic regression. It executes tasks to ensure proper formatting, quality, and suitability, laying the foundation for accurate analysis.

Here are the key tasks involved in the Data Preprocessing module:

1. **Missing Data Handling:** Addressing and managing missing values in the dataset.
2. **Feature Scaling:** Normalizing numerical features to ensure uniform impact during logistic regression.
3. **Categorical Encoding:** Converting categorical variables into a format suitable for analysis.
4. **Data Splitting:** Dividing the dataset into training and testing sets for model evaluation.

6.1.3 Logistic Regression Model Module:

The Logistic Regression Model employs logistic regression on the preprocessed data to predict thyroid conditions. It calculates probabilities, providing valuable insights for effective thyroid detection within the system.

Here are the key tasks involved in the Logistic Regression Model module:

1. **Model Training:** Utilizing the preprocessed data to train the logistic regression model.
2. **Prediction Generation:** Employing the trained model to generate predictions for thyroid conditions.
3. **Probability Estimation:** Calculating probabilities of thyroid disease presence based on input features.
4. **Model Evaluation:** Assessing the performance of the logistic regression model using metrics like accuracy and precision.
5. **Parameter Tuning:** Optimizing model parameters for enhanced predictive accuracy.

6.1.4 Prediction Visualization Module:

The Prediction Visualization Module interprets logistic regression outcomes, presenting them in a user-friendly format. It enhances understanding, providing clear visual insights into thyroid condition predictions for effective analysis and decision-making.

6.1.5 The Information Logging Module:

The Information Logging Module records essential details, including file information and logistic regression outcomes. This logging ensures comprehensive documentation for future reference, analysis, and continuous improvement, supporting the ongoing development and refinement of the thyroid detection system.

CHAPTER 7

PROJECT DESIGN

SL. No.	Attribute	Value Type	SL. No.	Attribute	Value Type
01	age	continuous	16	psych	f, t
02	sex	M, F	17	TSH measured	f, t
03	on thyroxine	f, t	18	TSH	continuous
04	query on thyroxine	f, t	19	T3 measured	f, t
05	on antithyroid medication	f, t	20	T3	continuous
06	sick	f, t	21	TT4 measured	f, t
07	pregnant	f, t	22	TT4	continuous
08	thyroid surgery	f, t	23	T4U measured	f, t
09	I131 treatment	f, t	24	T4U	continuous
10	query hypothyroid	f, t	25	FTI measured	f, t
11	query hyperthyroid	f, t	26	FTI	continuous
12	lithium	f, t	27	TBG measured	f, t
13	goitre	f, t	28	TBG	continuous
14	tumor	f, t	29	referral source	WEST, STMW, SVHC, SVI, SVHD, other
15	hypopituitary	f, t	30	category	Negative, hypothyroid, sick, hyperthyroid

```
Accuracy: 0.9373297002724795
Classification Report:
              precision    recall  f1-score   support

         0.0       0.95      0.99      0.97      1701
         1.0       0.68      0.27      0.39       134

    accuracy                           0.94      1835
   macro avg       0.81      0.63      0.68      1835
weighted avg       0.93      0.94      0.92      1835

Enter age: 20
Enter sex (1 for male, 0 for female): 0
Is the person sick? (1 for Yes, 0 for No): 1
Is the person pregnant? (1 for Yes, 0 for No): 0
Has the person undergone thyroid surgery? (1 for Yes, 0 for No): 0
Has there been a query on hypothyroidism? (1 for Yes, 0 for No): 1
Has there been a query on hyperthyroidism? (1 for Yes, 0 for No): 0
Does the person have goitre? (1 for Yes, 0 for No): 0
Does the person have a tumor? (1 for Yes, 0 for No): 0
Enter TSH level: 5.4
Enter TT4 level: 4.4
Enter T4U level: 6.7
Yes for thyroid
No for thyroid
No for thyroid
No for thyroid
Yes for thyroid
No for thyroid
No for thyroid
No for thyroid
No for thyroid
The model predicts that the person may have thyroid.
```

CHAPTER 8

CONCLUSION & FUTURE SCOPE

8.1 CONCLUSION:

- Logistic regression proves integral in thyroid detection, offering a robust and interpretable framework**.**
- Modules, from data import to prediction visualization, ensure efficient handling of CSV data and accurate modeling**.**
- The logging module supports continuous refinement for a dynamic and reliable thyroid detection system.
- Overall, logistic regression enhances diagnostic capabilities for thyroid conditions, advancing accuracy and accessibility**.**

8.2 FUTURE SCOPE:

- **Advanced Models**: Investigate the integration of more sophisticated machine learning models, like ensemble methods or deep learning, to elevate the precision and predictive power of thyroid disorder detection.
- **Real-time Monitoring:** Develop and implement real-time monitoring capabilities, enabling continuous tracking of thyroid conditions. This facilitates prompt intervention and ensures that the system remains responsive to dynamic health changes.
- **Enhanced User Interfaces:** Prioritize user interface enhancements to optimize user experience and facilitate seamless integration into clinical workflows. Intuitive interfaces empower healthcare professionals to efficiently interpret and act upon diagnostic results.
- **Personalized Medicine:** Embrace a personalized medicine approach by tailoring thyroid disorder predictions based on individual patient characteristics. This move towards precision medicine ensures treatment plans are customized, addressing the unique needs of each patient for improved healthcare outcomes.

APPENDIX (SAMPLE CODE)

```
# prompt: logistic regession thyroid for prediction
https://www.kaggle.com/datasets/emmanuelfwerr/thyroid-disease-data

from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
import numpy as np
import pandas as pd

# Load the data
df = pd.read_csv("/content/thyroidDF.csv")

df.columns

df.drop(['TSH_measured', 'T3_measured', 'TT4_measured', 'T4U_measured', 'FTI_measured',
'TBG_measured',
'patient_id', 'referral_source'], axis=1, inplace=True)
#mapping
diagnoses = {'-': 0, 'A': 1, 'B': 1, 'C': 1, 'D': 1,'E': 1, 'F': 1, 'G': 1, 'H': 1}

df['target'] = df['target'].map(diagnoses)


df.describe()

# ('age' > 100) to null
df[df.age > 100]
df['age'] = np.where((df.age > 100), np.nan, df.age)

# changed t ,f 0 and 1
df[['on_thyroxine', 'query_on_thyroxine' , 'on_antithyroid_meds', 'sick', 'pregnant', 'thyroid_surgery',
'I131_treatment', 'query_hypothyroid','query_hyperthyroid', 'lithium', 'goitre', 'tumor', 'hypopituitary',
'psych']] = df[['on_thyroxine', 'query_on_thyroxine' , 'on_antithyroid_meds', 'sick', 'pregnant',
'thyroid_surgery','I131_treatment', 'query_hypothyroid', 'query_hyperthyroid', 'lithium', 'goitre', 'tumor',
'hypopituitary', 'psych']].replace({'t': 1, 'f': 0})
# gender encode
df[['sex']] = df[['sex']].replace({'F': 1, 'M': 0})
df = df.fillna(value = 0)
feature_cols =['age', 'sex', 'on_thyroxine', 'query_on_thyroxine' , 'on_antithyroid_meds', 'sick', 'pregnant',
'thyroid_surgery', 'I131_treatment', 'query_hypothyroid',
# gender encode
df[['sex']] = df[['sex']].replace({'F': 1, 'M': 0})
df = df.fillna(value = 0)
```

```
feature_cols =['age', 'sex', 'on_thyroxine', 'query_on_thyroxine' , 'on_antithyroid_meds', 'sick', 'pregnant', 'thyroid_surgery', 'I131_treatment', 'query_hypothyroid',
# gender encode
df[['sex']] = df[['sex']].replace({'F': 1, 'M': 0})

df = df.fillna(value = 0)
feature_cols =['age', 'sex', 'on_thyroxine', 'query_on_thyroxine' , 'on_antithyroid_meds', 'sick', 'pregnant','thyroid_surgery', 'I131_treatment', 'query_hypothyroid', 'query_hyperthyroid', 'lithium', 'goitre', 'tumor','hypopituitary', 'psych', 'TSH', 'TT4', 'T4U', 'FTI']
X = df.loc[:, feature_cols]
y = df.target
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report
#  feature and target var
feature_cols = ['age', 'sex', 'on_thyroxine', 'query_on_thyroxine', 'on_antithyroid_meds', 'sick', 'pregnant','thyroid_surgery', 'I131_treatment', 'query_hypothyroid', 'query_hyperthyroid', 'lithium', 'goitre', 'tumor','hypopituitary', 'psych', 'TSH', 'TT4', 'T4U', 'FTI']
X = df[feature_cols]
y = df['target']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
logreg_model = LogisticRegression()
logreg_model.fit(X_train, y_train)
y_pred = logreg_model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred)
print(f'Accuracy: {accuracy}')
print('Classification Report:')
print(report)
user_input = {
    'age': float(input('Enter age: ')),
    'sex': int(input('Enter sex (1 for male, 0 for female): ')),
    'on_thyroxine': int(input('Is the person on thyroxine? (1 for Yes, 0 for No): ')),

  'query_on_thyroxine': int(input('Has there been a query on thyroxine? (1 for Yes, 0 for No): ')
'on_antithyroid_meds': int(input('Is the person on antithyroid medications? (1 for Yes, 0 for No): ')),
    'sick': int(input('Is the person sick? (1 for Yes, 0 for No): ')),
    'pregnant': int(input('Is the person pregnant? (1 for Yes, 0 for No): ')),
    'thyroid_surgery': int(input('Has the person undergone thyroid surgery? (1 for Yes, 0 for No): ')),
    'I131_treatment': int(input('Is the person on I131 treatment? (1 for Yes, 0 for No): ')),
'I131_treatment': int(input('Is the person on I131 treatment? (1 for Yes, 0 for No): ')),
    'query_hypothyroid': int(input('Has there been a query on hypothyroidism? (1 for Yes, 0 for No): ')),
    'query_hyperthyroid': int(input('Has there been a query on hyperthyroidism? (1 for Yes, 0 for No):
')),
```

```
    'goitre': int(input('Does the person have goitre? (1 for Yes, 0 for No): ')),
    'tumor': int(input('Does the person have a tumor? (1 for Yes, 0 for No): ')),
    'hypopituitary': int(input('Does the person have hypopituitary? (1 for Yes, 0 for No): ')),
    'psych': int(input('Is there a psychological issue? (1 for Yes, 0 for No): ')),
    'TSH': float(input('Enter TSH level: ')),
    'TT4': float(input('Enter TT4 level: ')),
    'T4U': float(input('Enter T4U level: ')),.
user_df = pd.DataFrame([user_input])
prediction = logreg_model.predict(user_df)
if prediction[0] ==1:
    print("The model predicts that the person may have thyroid.")
else:
    print("The model predicts that the person may not have thyroid.")
predicted_probabilities = logreg_model.predict_proba(X_test)[:, 1]
custom_threshold = 0.3
custom_predictions = (predicted_probabilities > custom_threshold).astype(int)
import pandas as pd
import numpy as np
y_pred = logreg_model.predict(X_test)
indices_with_ones = np.where(y_test == 1.0)[0]
X_test_ones = X_test.iloc[indices_with_ones]
predictions = logreg_model.predict(X_test_ones)
for prediction in predictions:
    if prediction == 1:
        print("Yes for thyroid")
    else:
        print("No for thyroid")
X_test_ones
     new_input = {
    'age': 73.0,
    'sex': 1.0,
    'on_antithyroid_meds': 0,
    'sick': 0,
    'pregnant': 0,
    'thyroid_surgery': 0,
    'I131_treatment': 0,
    'query_hypothyroid': 0,
    'query_hyperthyroid': 0,
    'lithium': 0,
    'goitre': 0,
    'tumor': 0,
    'hypopituitary': 0,
    'TSH': 47.0,
    'TT4': 52.0,
    'T4U': 0.90,
# Convert the input to a DataFrame
```

```
new_input_df = pd.DataFrame([new_input])
# Make a prediction
prediction = logreg_model.predict(new_input_df)
# Display the prediction
if prediction[0] == 1:
    print("The model predicts that the person may have thyroid.")
else:
    print("The model predicts that the person may not have thyroid.")
```

REFERENCES

1. Azar, a. T, Hassanien, A.E. and Kim, T. (2020), Expert system based on neural fuzzy rules for thyroid diseases diagnosis, Computer Science, Artificial Intelligence, ResearchGate, Pp. 1-12.

2. Anwar, S., Prasad, R., Chowdhary, B. S., et al. (2019), A telemedicine platform for disaster management and emergency care, Wireless Personal Communications, Springer, Pp. 191–204.

3. Biondi, B., G.J. Kahaly, and R.P. Robertson (2022), Thyroid Dysfunction and Diabetes Mellitus: Two Closely Associated Disorders, Reasearch Square, Pp. 789-824.

4. S. Sathya Priya, Dr. D. Anitha (2020), survey on "Thyroid Diagnosis using Data Mining Techniques", International Journal of Advanced Research in Computer and Communications Engineering,Research Gate, Pp. 4-7.

5. Milo T, Somech A(2020) Automating exploratory data analysis via machine learning: An overview. In: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, Springer ,pp.2617–2622.

6. Kumar, A., A.K. Tyagi, and S.K. Tyagi, (2021) data Mining: Various Issues and Challenges for Future A Short discussion on Data Mining issues for future work. International Journal of Emerging Technology and Advanced Engineering,Research Article, Pp.1157-1177.