THYROID DETECTION USING MACHINE LEARNING

Submitted By:

Nikhila Baby

MAC23MCA-2044

Faculty Guide:

Prof. Nisha Markose

Associate Professor

MCA Dept, MACE

INTRODUCTION

The thyroid gland has one of the most important functions in regulating metabolism. When the function of the thyroid gland is affected, it leads to inappropriate production of the thyroid hormone. Hypothyroidism and hyperthyroidism are two critical conditions caused by insufficient thyroid hormone production and excessive thyroid hormone production, respectively. The “Thyroid Detection Using Machine Learning” project is focussed on detecting and diagnosing thyroid disease.

The performance of three machine learning algorithms such as Random Forest, Logistic Regression, Support Vector Machine are compared to classify Thyroid disease into normal, hypothyroidism, or hyperthyroidism categories.

The dataset is taken from the Kaggle repository. The dataset contains 9172 sample observations and has 31 columns including 1 identifier, 1 class variable and 29 features.

The most significant features, which can be used to detect thyroid diseases more precisely are identified using forward feature selection, backward feature elimination, bidirectional feature elimination, and machine learning-based feature selection. The selected features are then used by the algorithms to build the models. Performance is evaluated and the best model is selected based on accuracy. Among the three algorithms, Random Forest is found to be best in terms of computational time and accuracy score, which make it significant for the proposed approach.

A blood test is one of the ways to diagnose thyroid disease. But after a lab blood test, a medical expert needs to examine the test stats of hormones and other parameters of the patient to diagnose the disease. There is very little difference in the blood test stats, which refer to different thyroid hormone levels. Such minor differences can lead to the wrong diagnosis even by medical experts as human error is expected. Incorrect diagnosis may lead to wrong medication and further complexities. So, an automated system can be very helpful to assist medical experts and even make automated disease predictions without any human mistakes.

LITERATURE SUMMARY

Page 1

Title of the paper	Chaganti, R., Rustam, F., De La Torre Díez, I., Mazón, J. L. V., Rodríguez, C. L., & Ashraf, I. (2022). Thyroid disease prediction using selective features and machine learning techniques. *Cancers*, *14*(16), 3914.
Area of work	Detection and classification of thyroid disease using selective features.
Dataset	Dataset was taken from UCI repository. The dataset contains 9172 sample observations and each sample is represented by 31 features.
Methodology / Strategy	The dataset consists of several thyroid-related disease records and many target classes. Four feature engineering approaches which includes forward feature selection (FFS), backward feature elimination (BFE), bidirectional feature elimination (BiDFE), and machine learning-based feature selection using an extra tree classifier are applied for feature selection. The aim is to classify thyroid disease to a multi-class problem.
Algorithm	RF, LR, SVM, ADA, GBM
Result/Accuracy	Results indicate that extra tree classifier-based selected features tend to provide the highest accuracy of 0.99 when used with the RF model. The lower computational complexity of the machine learning models like RF makes them good candidates for thyroid disease prediction. RF – 0.99 GBM – 0.98 ADA – 0.97 LR – 0.87 SVM – 0.92

Page 2

Title of the paper	Gupta, Punit, et al. "Detecting thyroid disease using optimized machine learning model based on differential evolution." *International Journal of Computational Intelligence Systems* 17.1 (2024): 3.
Area of work	Machine learning approach for thyroid disease detection using optimized model.
Dataset	The datasets used in this study are taken from the Kaggle repository. The thyroid disease dataset comprises 9172 samples and every sample has 31 features.
Methodology / Strategy	The dataset contains 25 target classes, of which the top 10 target classes are selected for experiments. The selected targeted dataset is imbalanced, so to make the dataset balanced, CTGAN augmentation technique is used. Train machine learning models with a training set and perform hyperparameter optimization using DE optimizer which helps to select the best hyperparameter setting for models.
Algorithm	RF, SVM, LR, AdaBoost, GBM
Result/Accuracy	AdaBoost with optimization shows an accuracy of 0.998. RF – 0.995 GBM – 0.996 AdaBoost – 0.998 LR – 0.643 SVM – 0.966

Page 3

Title of the paper	Hossain, M. B., Shama, A., Adhikary, A., Raha, A. D., Uddin, K. A., Hossain, M. A., ... & Bairagi, A. K. (2023). An explainable artificial intelligence framework for the predictive analysis of hypo and hyper thyroidism using machine learning algorithms. *Human-Centric Intelligent Systems*, *3*(3), 211-231.
Area of work	Classification of hypo and hyper thyroidism
Dataset	The data was taken from the UCI machine learning repository. Dataset contains 3221 instances with a total of 30 features.
Methodology / Strategy	The selected targeted dataset is imbalanced, so to make the dataset balanced, resampling techniques are used. Feature selection methods like univariate feature selection approach and the feature importance method is implemented and the best set of characteristics for building effective models are selected.
Algorithm	Decision Tree Classifer, Random Forest Classifer, Naive Bayes Classifer, Gradient Boosting Classifer, Logistic Regression Classifer, K- Nearest Neighbor, Support Vector Machine
Result/Accuracy	DT – 90.43 RF – 91.42 GBM – 90.5 NB – 67.86 KNN – 86.22 LR – 73.15 SVM – 73.7

PROJECT PROPOSAL

From the above three papers, we get to know that different models were used for the detection of thyroid disease. First paper is the detection and classification of thyroid disease using selective features. Second paper focuses on machine learning approach for thyroid disease detection using optimized model. Third paper aims at classification of hypo and hyper thyroidism.

The proposed system is the comparative study of three algorithms Random Forest, Logistic Regression and Support Vector Machine. The models will classify thyroid disease under three classes which are no thyroid, hyperthyroid and hypothyroid. An automated system can be very helpful to assist medical experts and even make automated disease predictions without any human mistakes. Patients can diagnose their condition without the assistance of a medical expert.

DATASET

The dataset is taken from the Kaggle repository. The dataset contains 9172 sample observations and has 31 columns including 1 identifier, 1 class variable and 29 features. The dataset contains numeric values and Boolean values. There are missing values in the dataset.

The identifier is the patient_id. The features are 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_measured, TSH, T3_measured, T3, TT4_measured, TT4, T4U_measured, T4U, FTI_measured, FTI, TBG_measured, TBG, and referral_source. The class labels include letters from A to T which indicates different thyroid conditions.

Dataset: https://www.kaggle.com/datasets/emmanuelfwerr/thyroid-disease-data