Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.14, May : 2024

AUTOMATED DIAGNOSIS TO PREDICT THE THYROID USING MACHINE LEARNING ALGORITHM

Mrs.Mrunal Pathak, Assistant Professor, Dept.Of Information Technology, Aissms IOIT Pune College, SPPU University.
Mrs.Meenakshi Thalor, Head of the Department, Dept. Of IT, SPPU University.
Siddhi Borse, Student at IT Dept, SPPU University.
Rutwik Khandgawhre, Student at IT Dept, SPPU University.
Aakanksha Kulkarni, Student at IT Dept, SPPU University.
Akanksha Raje, Student at IT Dept, SPPU University.

ABSTRACT
Contrary to Thyroid disorders affect millions of people worldwide, necessitating early and accurate diagnosis to ensure effective management and treatment. In this context, the Thyroid Stage Prediction App represents a groundbreaking healthcare innovation. This mobile application leverages cutting-edge machine learning technologies to predict thyroid stage progression with exceptional accuracy, offering a proactive approach to thyroid disease management. The Thyroid Stage Prediction App incorporates a user-friendly interface, enabling users to input their medical history, laboratory test results, and relevant symptoms. The app then processes this data using algorithms to assess the current thyroid stage and predict future developments. The predictive model is based on a vast dataset of anonymized patient records and is continually updated to ensure its reliability and precision. Key features of the app include stage-prediction, personalized recommendations, and alerts to consult with healthcare professionals, thus promoting early intervention and tailored treatment plans. Moreover, the app offers valuable educational content on thyroid health and wellness, empowering users to make informed decisions about their healthcare. The Thyroid Stage Prediction App promises to revolutionize thyroid disease management by promoting early detection, personalized care, and improved patient outcomes.

Keywords: Thyroid, Machine Learning, Logistic Regression, Support Vector Machine, Random Forest, Diagnosis Technique, Prediction.

I. Introduction
Thyroid disorders constitute a significant and widespread health concern, affecting millions of individuals globally [16]. Timely diagnosis and effective management of these conditions are paramount for ensuring the well-being of those affected. The Thyroid Stage Prediction Project introduces an innovative mobile application that harnesses the power of predictive technology to revolutionize the way we address thyroid-related health issues [5]. Thyroid diseases encompass a broad spectrum of conditions, from hyperthyroidism to hypothyroidism, and their complexities necessitate personalized approaches to diagnosis and treatment information. The Thyroid Stage Prediction App seeks to bridge this gap by utilizing advanced machine learning algorithms [1][7], drawing insights from extensive patient data, and offering an intuitive and user-friendly platform. In this era of digital health, our project aims to empower individuals by providing a tool that enhances early detection, promotes proactive intervention, and offers recommendations of hospitals and doctors for treatment. This introduction provides a glimpse of the trans-formative potential of our app, and the subsequent sections will delve deeper into its features, benefits, and the technology underpinning it [6].Thyroid disease has been the subject of a very small number of studies, and the authors have assessed many of these studies to provide a thorough foundation on the disease classification. Tahir Alyas proposed [1] Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach. This study used random forest and KNN algorithms to achieve an accuracy of 94.8%. Random forest is a machine learning algorithm that creates a multitude of

decision trees and then averages their predictions. KNN is a machine learning algorithm that classifies data points based on the similarity of their neighbor. The study found that random forest outperformed KNN in terms of accuracy [23]. The study also found that it is important to detect the stage of thyroid disease early, as failure in thyroid hormone production can be either excessive or inadequate. Early detection and treatment can help to prevent serious complications. Giuseppe Mollica proposed [2] Classification of Thyroid Diseases Using Machine Learning and Bayesian Graph Algorithms. This study used a Bayesian network framework to achieve good results in thyroid tumor classification. A Bayesian network is a type of probabilistic graphical model that represents the dependencies between variables. The study found that the Bayesian network framework was able to accurately classify thyroid tumors into different types.

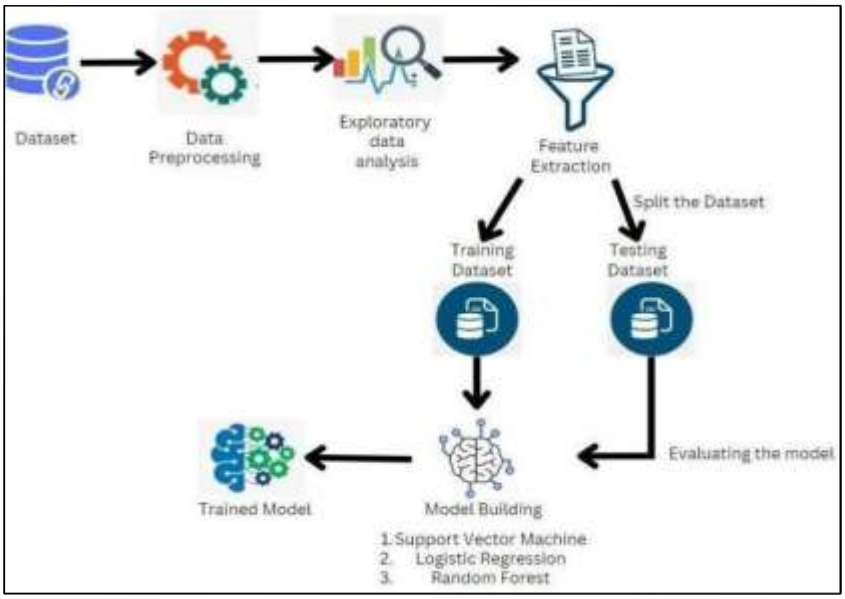

Figure 1: System Architecture of thyroid diagnosis project.

II. Methodology

In order to build a machine learning model, we need the data-set of thyroid patients. The Kaggle Machine Learning website provided the Thyroid data-set. we have considered a data set of 3772 patients. The data-set has various parameters and contains different values.
Used libraries:
Pandas: Used for data wrangling and data manipulation
NumPy: It is used in working with numerical values. Can mathematical operations easily. Sklearn: Implement machine learning model and statistical modelling. (Imported SVM, Logistic Regression, Random Forest)

2.1 Data Cleaning and Preprocessing

In our thyroid diagnosis project, data cleaning is a critical phase in the preprocessing stage. This involves identifying and correcting errors, inconsistencies, and inaccuracies in the dataset to bolster its reliability and quality. Implementing robust data cleaning techniques is essential for refining the dataset, ensuring more accurate predictions in thyroid diagnosis. This meticulous process not only eliminates noise but also establishes a solid foundation for subsequent stages, contributing to the overall robustness of our machine learning model.
Data Cleaning Techniques used in our project
1.Handling Missing Data: Delete rows or columns with a high proportion of missing values if it doesn't significantly impact the data-set. Imputation: Fill in missing values using techniques such as mean.

2.Handling Outliers: Identify outliers using statistical methods like z-scores or the IQR (Inter-quartile Range) and consider whether they should be removed or transformed.
3.Transformation: Apply mathematical transformations like logarithm, square root, or Box-Cox to mitigate the impact of outliers.
4.Handling Inconsistent Data: Correct inconsistent data entries by standardizing formats, fixing typos, and reconciling discrepancies in naming conventions.
5.Removing Irrelevant Data: Eliminate columns that do not contribute meaningful information to the problem.

2.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an essential step in the data analysis process that involves examining and visualizing the data set to gain insights and identify patterns, relationships, and anomalies. It gives you a basic understanding of the data and its distribution. We used Count-plot, Hist plot, Pie plot, Pandas Profiling plots to print data visually.
In order to understand the dataset better we used the pandas profiling library to get insight into the dataset. It uncovered the relationship between different parameters of the dataset. Handling Imbalanced data in the dataset.
We handled imbalanced data in the dataset by over-sampling technique. Over-sampling is used when the amount of data collected is insufficient. A popular over-sampling technique is SMOTE (Synthetic Minority Over-sampling Technique), SMOTE creates synthetic samples by randomly sampling the characteristics from occurrences in the minority class. We did it to improve the accuracy of the result [16].
Shape before the Oversampling: (2896, 14)
Shape after the Oversampling : (5340, 14)

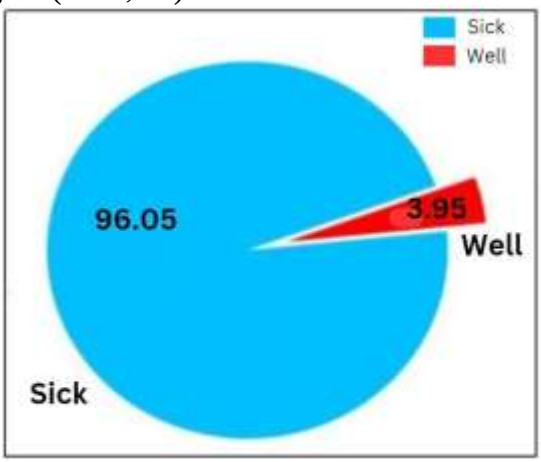

Figure 2: Pie Plot of sick and well people ratio in dataset.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.14, May : 2024

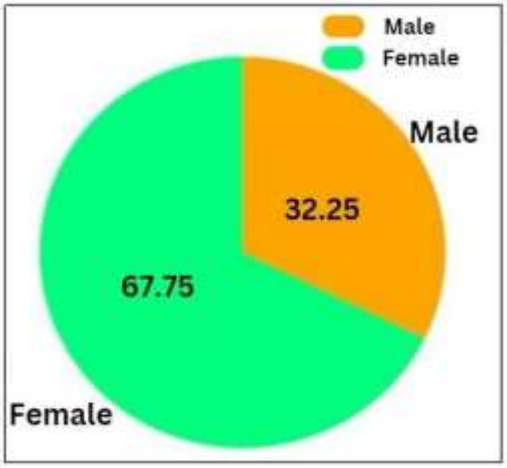

Figure 3: Pie Plot of male and female ratio in dataset

2.3 Feature Extractions

Feature extraction is the process of transforming raw data into numerical features that can be processed while preserving the information in the original data set. This is particularly important in machine learning when you're dealing with high-dimensional data or when you want to enhance the performance of your models [9].

After feature extraction we got these features. These features will be used in model building to get exact result and high accuracy. Feature extraction is the most significant step in any machine learning project.[14][15]

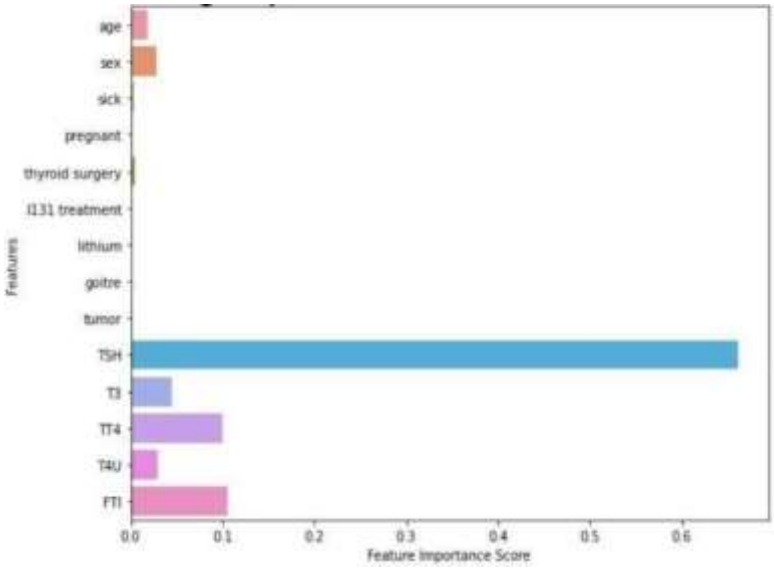

Figure 4: Feature's importance

Table 1: Extracted Features

	Extracted Feature	Value Type
1	Gender	T/F
2	Age	continuous

3	Pregnant	T/F
4	Sick	T/F
5	Thyroid Surgery	T/F
6	Goiter	T/F
7	Tumor	T/F
8	TSH	continous
9	T3	continous
10	TT4	continous
11	T4U	continous
12	FTI	continous
13	Lithium	T/F

Transforming non numerical labels to numerical:
We converted non numerical labels to numerical so it would be easy for model building. All the attributes having values as true or false got converted to 0 and 1. Such attributes are gender, pregnant, sick, tumor, goiter, thyroid surgery, lithium, T131 treatment.
Normalization by scaling: The values of the features in dataset are on different scale so it reduces the accuracy of the model. We used scaling for TT4, FTI and age. Normalization helps to remove the impact of the scale and put all features on the same scale

Pseudo Code :

```
    models = {
  LogisticRegression(max_iter=500):'Logistic Regression',
  SVC():"Support Vector Machine",
  RandomForestClassifier():'Random Forest'
}
for m in models.keys():
  m.fit(x_smote,y_smote)
for model,name in models.items():
   print(f"Accuracy Score for {name} is : ",model.score(X_test,y_test)*100,"%")
```

III. Results

In order Splitting data into appropriate subsets is a fundamental step in preparing a dataset for machine learning. The primary goal is to have separate sets for training, validation, and testing. A common approach is the 70/30 or 80/20 split, where the majority of the data is used for training and the rest for validation and testing. We have divided the dataset into 80/20 ratios for training and testing respectively.

Logistic Regression

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.14, May : 2024

Logistic Regression is a statistical and machine learning model used for binary classification tasks, as our goal was to predict one of two possible outcomes. It's a simple yet powerful algorithm that's widely used in various fields. So, we used logistic regression to find a s-curved line on a graph dividing true and false values in two different sections. By considering attributes such as gender, age, pregnant, sick, Thyroid surgery, goiture, tumor, TSH, T3, TT4, T4U, FTI, Lithium, I131 Treatment This is used to find whether patients have thyroid or not [4].

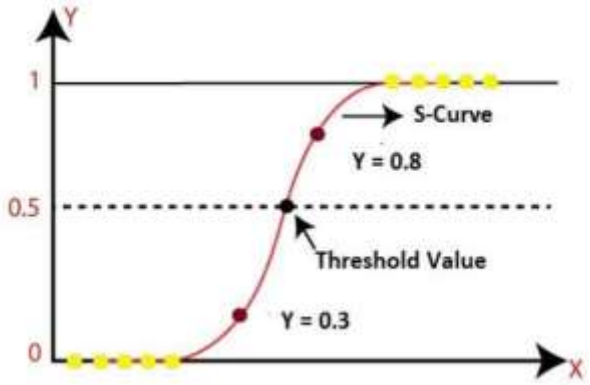

Figure 5: Graph of logistic regression

Support Vector Machine

A Support Vector Machine (SVM) is a powerful supervised machine learning algorithm for classification and regression tasks [15]. It's especially popular in binary classification problems but can be extended to multi- class classification as well. Our main idea behind using SVM is to find a hyper plane by using extracted features such as gender, age, pregnant, sick, Thyroid surgery, goiture, tumor, TSH, T3, TT4, T4U, FTI, Lithium, I131 Treatment. After 500 iterations we were able to find the maximum margin between the those features and generated the result showing whether the patient has thyroid or not [11].

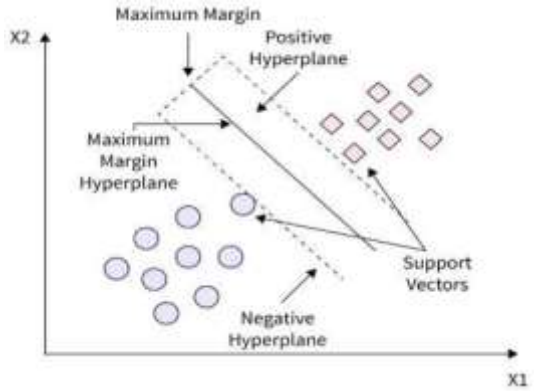

Figure 6: Algorithm and Graph of Support Vector Machine

Random Forest

Random Forest is a popular ensemble learning method used in machine learning for both classification and regression tasks. It is based on the idea of creating multiple decision trees during training and combining their predictions to improve accuracy [3] and reduce over fitting. We used this technique to improve the accuracy of the model. All the attributes from the feature extraction were used to build various decision trees and at last, all those were combined to find the exact output.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.14, May : 2024

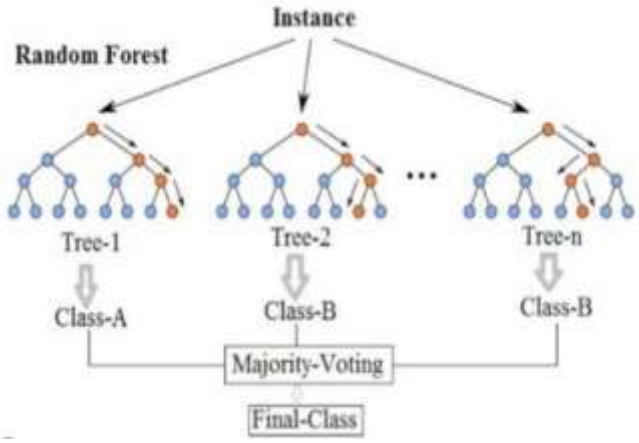

Figure 7: Random Forest feature analysis and classification in classes

Model Evaluation

At last evaluations were done of all the models to find out the model's accuracy. To find out the accuracy we used following techniques which are support vector machine, logistic regression, random forest.

Classification Report for Support Vector Machine

	precision	recall	f1-score	support
0	0.82	0.98	0.89	54
1	1.00	0.98	0.99	670
accuracy			0.98	724
macro avg	0.91	0.98	0.94	724
weighted avg	0.98	0.98	0.98	724

Classification Report for Logistic Regression

	precision	recall	f1-score	support
0	0.86	0.91	0.88	54
1	0.99	0.99	0.99	670
accuracy			0.98	724
macro avg	0.93	0.95	0.94	724
weighted avg	0.98	0.98	0.98	724

Classification Report for Random Forest

	precision	recall	f1-score	support
0	0.93	0.93	0.93	54
1	0.99	0.99	0.99	670
accuracy			0.99	724
macro avg	0.96	0.96	0.96	724
weighted avg	0.99	0.99	0.99	724

References

[1] T. Alyas, M. Hamid, K. Alissa, T. Faiz, N. Tabassum, and A. Ahmad, "Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach," 2022. [Online]. Available: https://doi.org/10.1155/2022/9809932.

[2] G. Mollica, D. Francesconi, G. Costante, S. Moretti, R. Giannini, E. Puxeddu, P. Valigi, "Classification of Thyroid Diseases Using Machine Learning and Bayesian Graph Algorithms," IFAC PapersOnLine, vol. 55-40, pp. 67–72, 2022.doi :10.1016/j.ifacol.2023.01.0502

[3] R. Jha, V. Bhattacharjee, A. Mustaf, "Increasing the Prediction Accuracy for Thyroid Disease: A Step Towards Better Health for Society," 2021. [Online]. Available: https://doi.org/10.1007/s11277-021-08974-3.

[4] P. Isawasan, "Regression Study for Thyroid Disease Prediction: Comparison of Crossing-Over Approaches and Multivariate Analysis," 2022.

[5] M. Riajuliislam, K. Zahidur Rahim, A. Mahmud, "Prediction Of Thyroid Disease (Hypothyroid) In Early Stage Using Feature Selection And Classification Techniques," 2021.DOI:10.1109/ICICT4SD50815.2021.9397052

[6] L. Aversano, "Thyroid Disease Treatment prediction with machine learning approaches," *Procedia Computer Science*, vol. 192, pp. 1031–1040, 2021. [Online]. Available: https://doi.org/10.1016/j.procs.2021.08.106

[7] M. R. Islam, K. Z. Rahim, and A. Mahmud, "Prediction of Thyroid Disease (Hypothyroid) in Early Stage Using Feature Selection and Classification Techniques," in *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, 2021, pp. 60-64. [Online]. Available: https://dx.doi.org/10.1109/ICICT4SD50815.2021.9397052

[8] A. K. P and J. V. B. Benifa, "A comprehensive analysis using neural network-based model for thyroid disease prediction," in *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, 2022, pp. 72-78.

[9] R. Chaganti, F. Rustam, I. De La Torre Díez, J. L. V. Mazón, C. L. Rodríguez, and I. Ashraf, "Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques," *Cancers*, vol. 14, no. 16, p. 3914, 2022. [Online]. Available: https://doi.org/10.3390/cancers14163914

[10] G. Chaubey, D. Bisen, S. Arjaria, and V. Yadav, "Thyroid disease prediction using machine learning approaches," *National Academy Science Letters*, vol. 3, pp. 128–133, 2021. [Online]. Available: http://dx.doi.org/10.1007/s40009-020-00979-z

[11] R. Banu, "Classification model using random forest and SVM to predict thyroid disease," *International Journal Of Scientific & Technology Research*, vol. 9, no. 2, pp. 1680–1685, 2018.

[12] L. Aversano, M. L. Bernardi, M. Cimitile et al., "Thyroid disease treatment prediction with machine learning approaches," *Procedia Computer Science*, vol. 192, pp. 1031–1040, 2021.

[13] A. R. Rao and B. S. Renuka, "A machine learning approach to predict thyroid disease at early stages of diagnosis," in *2020 IEEE International Conference for Innovation in Technology (INOCON)*, Bengaluru, India, 2020, pp. 2020–2023. [Online]. Available: https://doi.org/10.1109/INOCON50539.2020.9298252

[14] A. Aswathi and A. Antony, "An intelligent system for thyroid disease classification and diagnosis," in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Coimbatore, India, 2018, pp. 1261–1264. [Online]. Available: https://doi.org/10.1109/ICICCT.2018.8473349

[15] K. Shankar, S. Lakshmana Prabu, D. Gupta, A. Maseleno, and V. Albuquerque, "Optimal feature-based multi-kernel SVM approach for thyroid disease classification," *The Journal of Supercomputing*, vol. 28, no. 76, pp. 1128–1143, 2020. [Online]. Available: https://doi.org/10.1007/s11227-022-04941-2

[16] A. K. Aswathi, and A. Antony, "An Intelligent System for Thyroid Disease Classification and Diagnosis," in *2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018)*, 2018, pp. 1261–1264. [Online]. Available: https://doi.org/10.1109/ICICCT.2018.8473349