***Abstract* — Thyroid disease encompasses a broad spectrum of disorders ranging from simple to malignant tumors, presenting a significant challenge in diagnosis and management.Traditional approaches rely heavily on clinical examination, laboratory tests, and imaging studies leading to delays. Machine Learning techniques have emerged as powerful tools in healthcare, offering promising avenues for improving accuracy and efficiency.**

**This review provides an overview of the ML-driven approaches to enhance clinical decision-making in healthcare settings.The dataset used for training and testing the model comprises clinical features, laboratory results and findings from patients with confirmed thyroid disease diagnoses.**

**Overall, this study showcases the efficiency of Random Forest algorithm in thyroid disease diagnosis and underscores**

**Its potential to serve as a valuable decision-support tool.**

I. INTRODUCTION

Computational biology has significantly advanced healthcare by utilizing stored patient data to predict diseases.With the aid of prediction algorithms,early-stage disease diagnosis has become possible. While medical information systems are abundant in datasets, there is a scarcity of intelligent systems capable of effectively analyzing diseases Over time, machine learning algorithms have become instrumental in addressing complex and nonlinear problems in model development.

In disease prediction models, it's crucial to carefully select features from diverse datasets to ensure accurate classification of healthy patients. Failure to do so can result in misclassification, potentially leading to unnecessary treatment for healthy individuals.

The thyroid gland, located in the human neck beneath Adam's apple, is an essential endocrine gland responsible for secreting thyroid hormones, namely levothyroxine (T4) and triiodothyronine (T3). These hormones play a pivotal role in regulating metabolism, protein synthesis, heart rate, calorie burning, body temperature, energy levels, and protein management. Iodine serves as the primary component of the thyroid gland, and its deficiency can lead to various thyroid-related disorders, such as hyperthyroidism and hypothyroidism.

Several factors contribute to thyroid problems, including inadequate hormone supply, ionizing radiation exposure, chronic thyroid inflammation, iodine deficiency, and enzyme deficiencies impairing thyroid hormone production. Various treatment options, including medication and thyroid surgery, are available for managing thyroid disorders.

II. DEFINITION OF PROBLEM

Statistics show that thyroid diseases are becoming more common in India. Approximately one in every ten Indian adults has thyroid disorders. It is estimated that over 42 million people suffer from thyroid illness. Predicting thyroid disorders by a doctor is a time-consuming process that may result in negative predictions; only experienced doctors can thoroughly analyze the case. Machine learning can help doctors diagnose diseases and lessen their impact.

**A. Objective**

The main objective of this project is to predict if a person is having compensated hypothyroid, primary hypothyroid, secondary hypothyroid, hyperthyriod or negative (no thyroid) with the help of Machine Learning. The main objective is to develop a system that can predict the type of thyroid disease that a patient is affected by. To predict all possible types of Thyroid diseases.

III. LITERATURE SURVEY

# [1] XGBoost Multi-class Classification By [Emmanuel F. Werr,](https://www.kaggle.com/emmanuelfwerr) Kaggle Expert, Barcelona, Catalonia, Spain, MSc Student. In this notebook, they have cleaned their dataset and used the XG-Boost algorithm. They have analyzed the accuracy of different xg-boost algorithms used and comparison is made to find the best one with high accuracy.

[2]Kwang-Sig Lee, Hyuntae Park, AI Center, Korea University College of Medicine, 02841 Seoul, Republic of Korea.

# In this research review, their study tells us that random forest variable importance would vary across different types of data for the early diagnosis of thyroid disease.The authors evaluate the performance of various machine learning methods, noting that the random forest and gradient boosting are effective for numeric data.This literature survey serves as a foundation for future research endeavors, guiding researchers toward addressing existing challenges and unlocking the full potential of machine learning in advancing thyroid disease diagnostics.

[3] Ankita Tyagi and Rikitha Mehra “Interactive Thyroid Disease Prediction System Using Machine Learning Techniques'' 5th IEEE International Conference on Parallel, Distributed and Grid Computing(PDGC-2018), 20-22 Dec, 2018, Solan, India.

In this work, they use different classification algorithms- Decision Tree, Support Vector Machine, Artificial Neural Network, k-nearest-neighbor algorithm. Based on the data set obtained from the UCI Repository, classification and prediction was performed and accuracy was obtained based on output produced. They have analyzed the accuracy of algorithms used and comparison is made to find the best technique with high accuracy.

[4] Hitesh Garg. Feed Forward Neural Network is used for feature extraction and segmentation from Ultrasound images to predict the tumors. The accuracy and other factors were measured and all the average values were above 86%.

IV. SYSTEM ANALYSIS AND DESIGN

A. METHODOLOGY

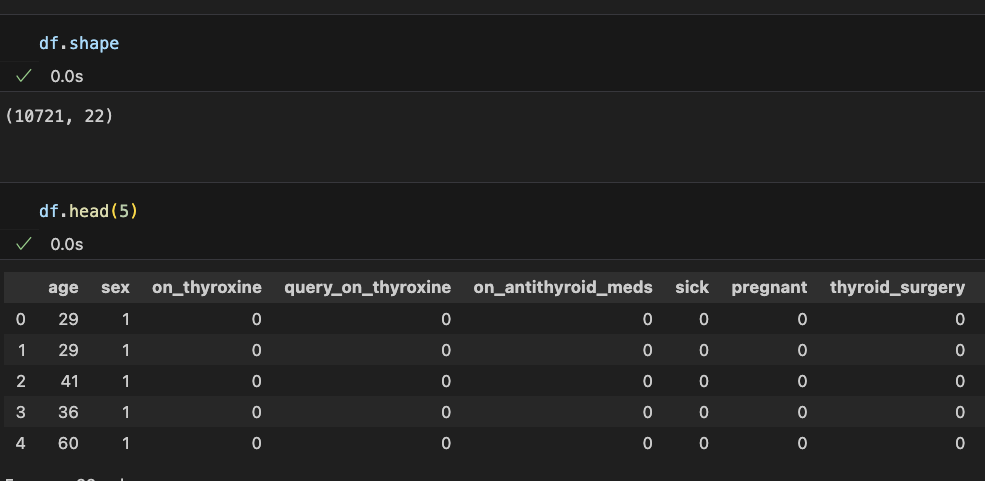
To examine and predict Thyroid disease, a blood report is required. The thyroid blood test data set will be analyzed using several supervised machine learning classifier algorithms. Based on the accuracy of several algorithms, the best accuracy algorithm will be picked to obtain the outcome.

For the first part, the thyroid data set is taken from the UCI repository and Kaggle.The UCI repository dataset has types of hypothyroidism and Kaggle has hyperthyroidism. Two datasets are combined to get the resultant dataset. The dataset of hyperthyroidism and hypothyroidism is used where hyper, primary- hypo, secondary-hypo, compensated hypo are the different labels. These data sets need to be checked before feeding it to training. There may be presence of null data or unnecessary data, this should undergo data cleaning to remove such data. Cleaned data is used as training data and test data, which is fed as input to the algorithm.

The algorithm extracts the features from different dataset to classify the data according to the labels. To check the accuracy of the prediction, test data is fed to the algorithm. Based on the feature extracted, probability will be generated for test data by comparing the features of both. Highest probability value will be classified to that particular label whether it is primary hypothyroidism, secondary hypothyroidism, compensated hypothyroidism or hyperthyroidism.

Web app is developed using Express JS and Node JS and the web application is designed using HTML5 ( front end), CSS, JS, and where the chosen ML model (using flask) will be linked with web app and HTML. MongoDB database is used for login and signup systems. The user's blood test data will be entered in the web app front end and the back end will process the data using the model and the result will be designed on screen.

Dataset newboth.csv has 10,721 samples and 21 features. 5 lab test results are used to try to predict whether a patient’s thyroid to the class is negative, primary hypothyroidism, secondary hyperthyroidism, compensated hyperthyroidism or hyperthyroidism. The diagnosis (the class level) was based on a complete medical record. This dataset has 10248 instances of negative class, 183 instances of hyperthyroidism class, 2 instances of secondary hypothyroidism class, 184 instances of compensated hypothyroidism and 95 instances of primary hypothyroidism class.



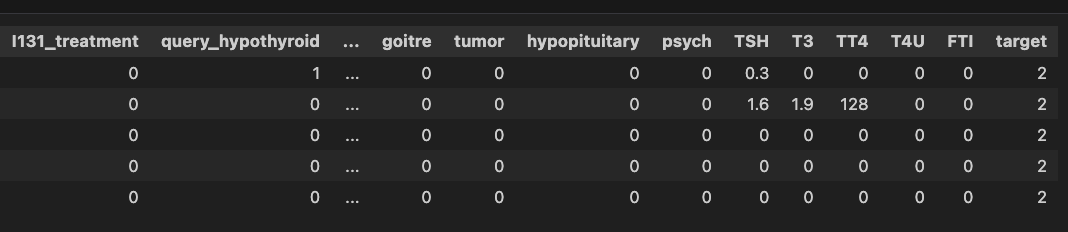


Fig 2: dataset header

V. EVALUATION AND TESTING

A. Result

* Heat map

A heat map (or heatmap) is a data visualization technique that shows the magnitude of a phenomenon as color in two dimensions. The color variation may be by hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space. Correlation between parameters of our dataset is interpreted and pictorial view is obtained as shown in fig 3.

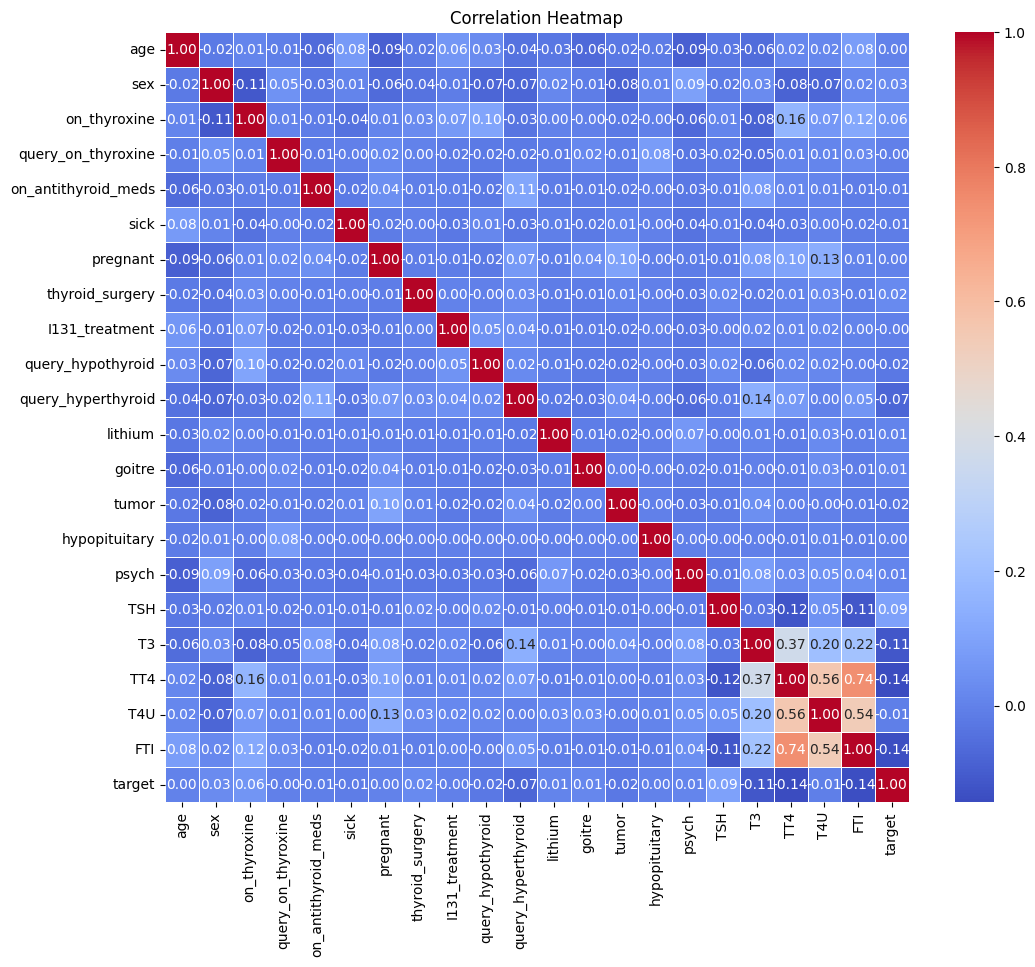


Fig 3: Heat Map

Random forest classifier algorithm has been implemented and 98.04% accuracy score is obtained as shown in fig 4.

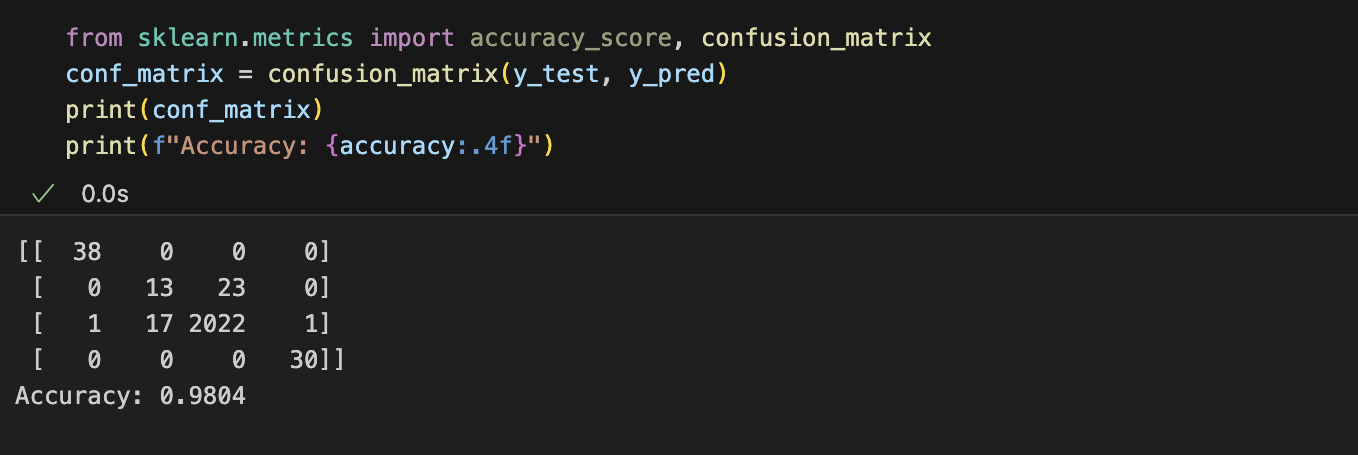


Fig 4: Random forest algorithm accuracy

The XGBoost algorithm has been implemented and a 98% accuracy score is obtained as shown in fig 5.

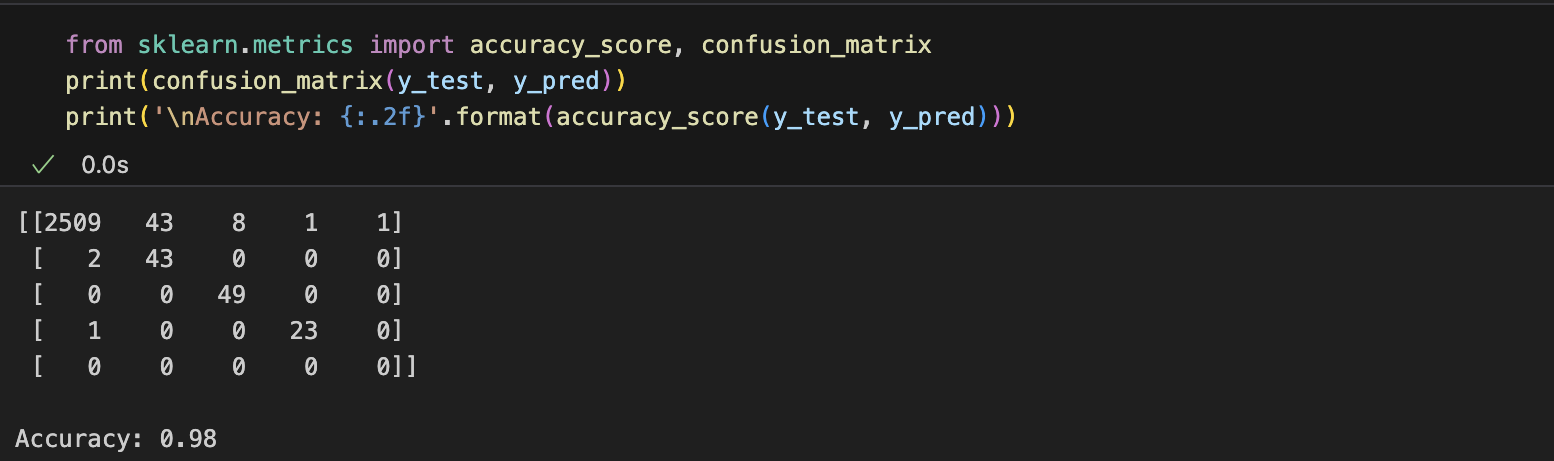


Fig 5: XGBoost algorithm accuracy

KNN algorithm has been implemented and a 97.02% accuracy score is obtained as shown in fig 6.

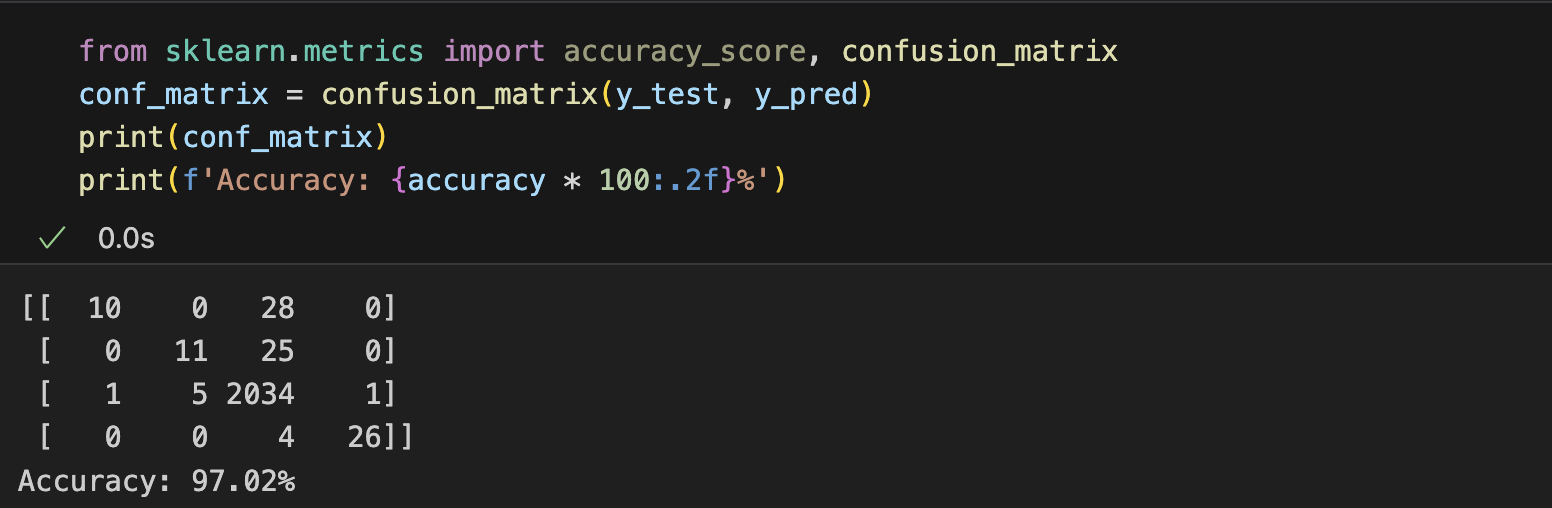


Fig 6: KNN algorithm accuracy

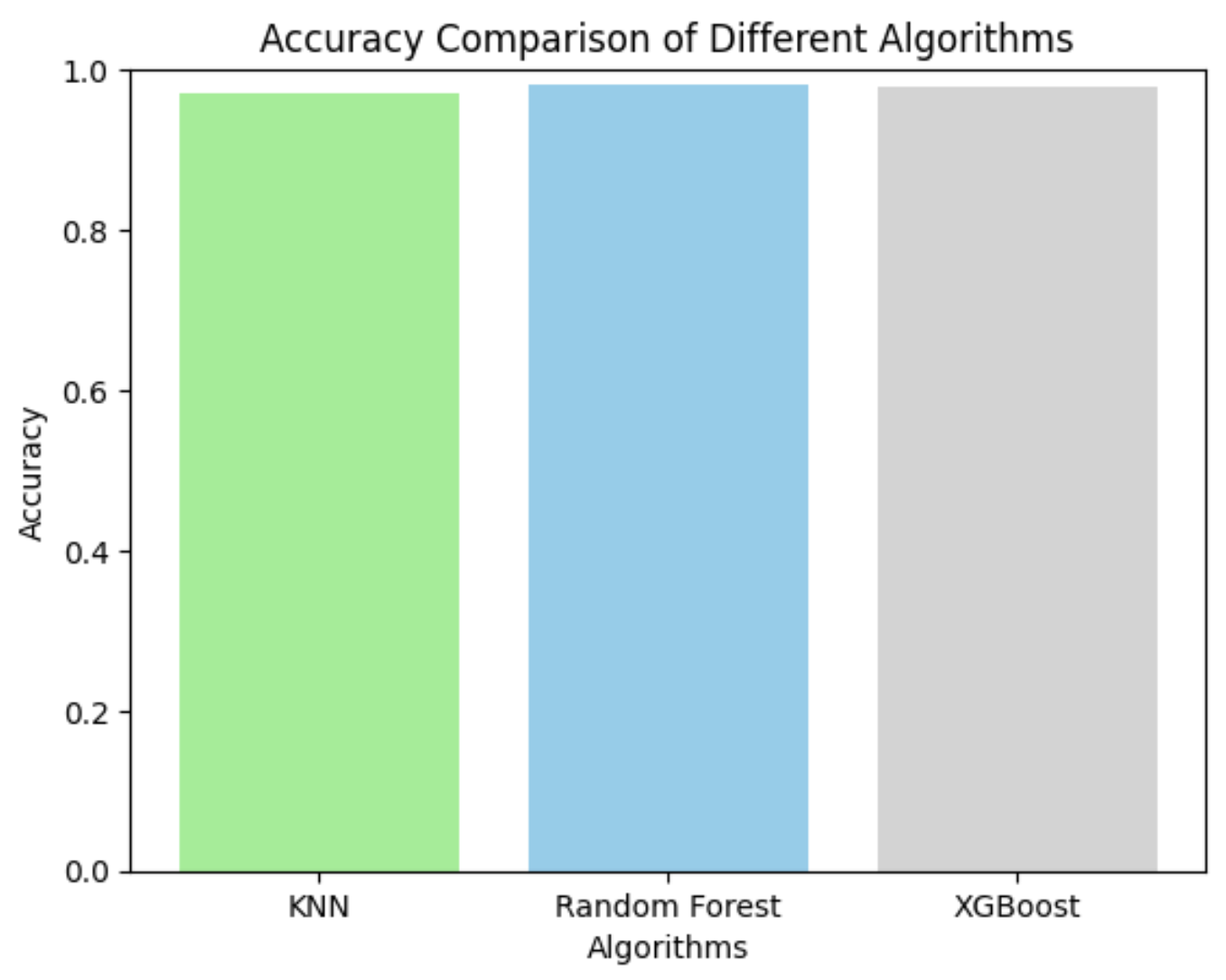


Fig 9: bar graph of score

Since accuracy obtained from the random forest classifier model (98.02%) was highest, this model will be considered for our prediction model. To save this model for interfacing with web apps, the joblib library is used and the file “newboth.pkl” will be created.

Web apps are created to interface users and the trained model. Python Flask coding is used to create web apps and html is used to design web pages.

Fig 10 shows the data entered into the parameter box of the web app created, which will be used at the back end by the model to predict thyroid disease.

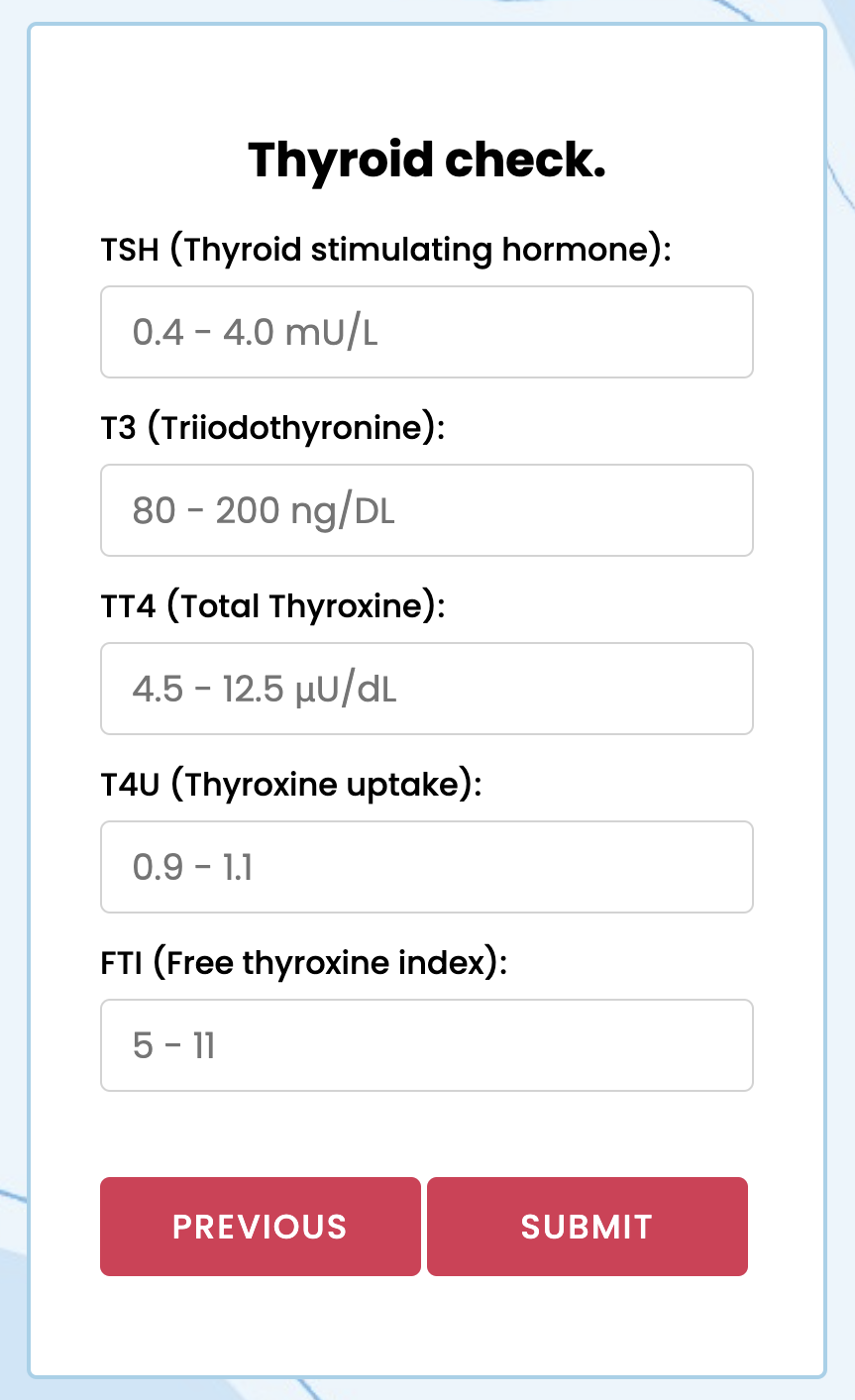


Fig 10: prediction example

For the data given through the web page as shown in fig 10, after processing the data by the model, the result will be displayed in the same web page as shown in fig 11.

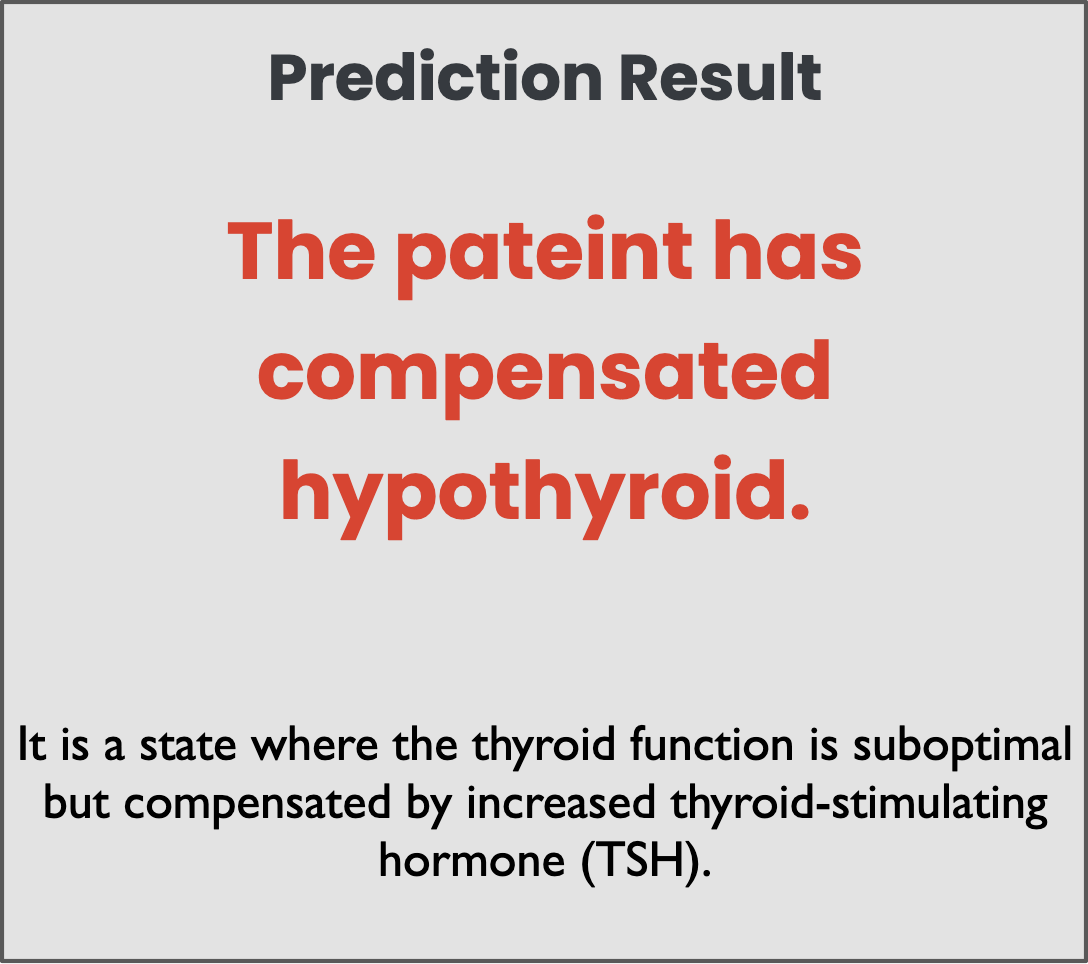


Fig 11: result page

VI. CONCLUSION

Thyroid Detection using Machine Learning is a project idea that aims at a smart and precise way to predict thyroid disease. We have made use of the Random Forest Classifier algorithm to train our dataset and to predict thyroid disease with more accuracy. Here the machine is trained to detect whether the person is normal, compensated hypothyroid, primary hypothyroid, secondary hypothyroid or hyperthyroidism based on the user’s input. So when a user enters data in a web app the data will be processed in the backend (model) and the result will be displayed on the screen. Our objective was to give society an efficient and precise way of machine learning that can be used in applications aiming to perform disease detection.

VII. REFERENCES

[1][Emmanuel F. Werr,](https://www.kaggle.com/emmanuelfwerr) Kaggle Expert. Link: <https://www.kaggle.com/code/emmanuelfwerr/xgboost-multi-class-classification/notebook>

[2]Kwang-Sig Lee, Hyuntae Park, AI Center, Korea University College of Medicine, 02841 Seoul, Republic of Korea. Links: <https://doi.org/10.31083/j.fbl2703101>

https://article.imrpress.com/journal/FBL/27/3/10.31083/j.fbl2703101/2768-6698-27-3-101.pdf

[3] Ankita Tyagi and Ritika Mehra. (2018).“Interactive Thyroid Disease Prediction System using Machine Learning Techniques” published on ResearchGate.

[4] Hitesh Garg,(2013). “Segmentation of Thyroid Gland in Ultrasound image using Neural Network” published on IEEE.

* https://www.kaggle.com/code/emmanuelfwerr/xgboost-multi-class-classification/notebook
* https://archive.ics.uci.edu/dataset/102/thyroid+disease