iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** I **Month of publication:** January 2023

DOI: https://doi.org/10.22214/ijraset.2023.48531

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue I Jan 2023- Available at www.ijraset.com

Thyroid Disease Detection System

Devansh Sirohi[1], Deepanshu Kashyap[2], Devendra Pal[3], Gopal Goyal[4], Bhumica Verma[5]
[1, 2, 3, 4, 5] *IMSEC Ghaziabad*

Abstract: ***This paper aims to recognize the diagnosis of the thyroid disease and then categorize the type of thyroid disease a patient may be suffering from (i.e., hyperthyroidism or hypothyroidism). The project implementation is being done by using python and Kaggle is the platform from which the dataset has been taken. At present many machine learning algorithms have been used to detect thyroid disease like but our goal is to implement the machine learning algorithm which has higher accuracy and which takes less time in detecting the disease along with the type of thyroid. We have trained the dataset taken from Kaggle over various machine learning techniques. We have also attempted to reduce the number of parameters required to detect the disease.***

I. INTRODUCTION

As per statistics data, the thyroid disease is a major disorder nowadays. Out of every ten adults one adult suffer from thyroid. According to an estimation above forty million people suffer from this problem of thyroid. Detection of thyroid disease using traditional process is a tedious process. In some case it has been seen that sometimes traditional process leads to wrong prediction of result due to less preciseness using traditional process. This system can predict the accurate result by using a machine learning algorithm known as Random Forest which is providing more accuracy as compared to more such algorithms used for the same parameters.

The diagnosis of thyroid disease is a tedious process which requires efforts and time to dictate the results through traditional process done by doctors. Sometimes even traditional process leads to false results. This process can be made effortless and precise by applying some machine learning techniques to it. This system will automate the process of diagnosis of thyroid disease by taking few parameters and will dictate an accurate and efficient result in less time.

Due to the less precision in traditional diagnosis process and sometimes false prediction by the doctors the patient has worse impact over their health. To improve this and to make doctor's work effortless this system can be used which will help them in diagnosis at an early level so that the further treatment can be carried out. This system will make this process easier and very fast. It will also be more accurate and efficient.

Our objectives in this study is to increase the accuracy of prediction of result, attempt to make the tedious process of prediction easy for doctors and clearly dictate the result by specifying the type of thyroid disease.

A person who is getting weak excessively or vice-versa can get the diagnosis done anytime as the system will be deployed as an app or a website. This service will be taken online as compared to the offline traditional process. Users will not be charged any fee for using the system which will make the thyroid diagnosis a free of cost service. One the basis of diagnosis result the user can reach the doctor for further treatment and medications.

II. LITERATURE SURVEY

In Literature, many studies had been conducted and they dealing with detection of Thyroid disease by the use of hormonal parameters and personal data of patient such as T3 and T4 levels. The most common technology used is machine learning, prediction models and deep neural network model.

Authors also focuses on various machine learning algorithms such as Support Vector Machine **(**SVM**),** Decision Tree, Gradient Boosting for performing diagnosis of thyroid disease. Their study shows that the Decision Tree has the accuracy of 99.23% and this can be used successfully to detect thyroid disease.

These authors conducted studies to accomplish their goal of predicting thyroid diseases using many data mining methods and finding relation between different components like T3, T4, TSH levels and their relation with hyperthyroidism and hypothyroidism These authors have taken their study of Support Vector Machine, Decision Tree and various data mining algorithm, apply these algorithms on datasets to verify their results. Authors have conducted studies to diagnosing the two most common thyroid diseases i.e., Hyperthyroidism and Hypothyroidism.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue I Jan 2023- Available at www.ijraset.com

This study is accomplished using mainly two methods like Logistic Regression and Neural network algorithm. The study has been conducted on 435 group of people and the model took hormonal and physical parameters as input. The model shows 95.5% accuracy with neural algorithms and 92.3% accuracy with Logistic Regression in most of the cases.
This study differs from the previous study which is been made using decision tree as it has a higher accuracy than these models .it is the first model which used neural network algorithms like PNN, GRNN and this study provide very short, precise and crisp results. The main objective of this work is to use all data we have gathered so far on different patients to predict LT4 - based treatments needed to be used or not. According to our goal i.e., detection of thyroid disease across the all the study we have done, we use and do the comparison a of all different machine learning and neural network algorithms to predict whether a patient is suffering from thyroid disease.
Finally, there is one more contribution of work we have done is represented by the dataset we have extracting by obtaining real time data of multiple patients. This dataset is formed by integration of two datasets which are of patient's current state and patient's previous medical history and treatment.

III. PROPOSED SYSTEM

In the approach, we take a dataset of the thyroid disease from the Kaggle website and performing the different process of the preprocessing to make the dataset suitable for the performing the different machine learning algorithms as like Logistic Regression, Random Forest, Gradient Boosting etc.
There are several algorithms which can be applied on the dataset but we choose the best algorithm, we produce the more accurate result in very efficient time period.
After preforming the algorithms of the machine learning and analyzing the results of the different algorithms we finally choose the best accurate algorithm which is finally used in our model.

A. Input Dataset

This directory contains the latest version of an archive of thyroid diagnoses obtained from the Garvan Institute, consisting of 9172 records from 1984 to early 1987. Each record looks like
1) (29 attribute values), diagnoses [record identification].

B. Preprocessing

In this step of the process, we make the row dataset suitable for a building and training Machine Learning models.

C. Feature Selection

In this step we select the important features(columns) from the row dataset. In this study we take the 17 features in which 15 features are work as the input which are given to the model and 2 features work as the output for analyzing the result.

D. Splitting Dataset

Now, we split the dataset in to the training and testing data with the ratio of the 70% and 30% of the original dataset. Now dataset is ready to work upon the different machine learning algorithms.

E. Logistic Regression

Logistic regression is a statistical method used for classification tasks. It is a supervised learning algorithm that takes a set of input features and uses them to predict a binary outcome, such as "yes" or "no," "true" or "false," or "0" or "1." In logistic regression, the model learns the relationship between the input features and the binary outcome by fitting a logistic curve to the data. The logistic curve is a sigmoid function that maps any input value to a value between 0 and 1. This output value can be interpreted as the probability that the binary outcome will be $^{11}1,^{11}$ given the input features. To make a prediction, the logistic regression model uses the learned logistic curve to calculate the probability of the binary outcome being $^{11}1,^{11}$ given a new set of input features. If the probability is above a certain threshold (usually 0.5), the model predicts $^{11}1,^{11}$ otherwise it predicts $^{11}0.^{11}$

©IJRASET: All Rights are Reserved

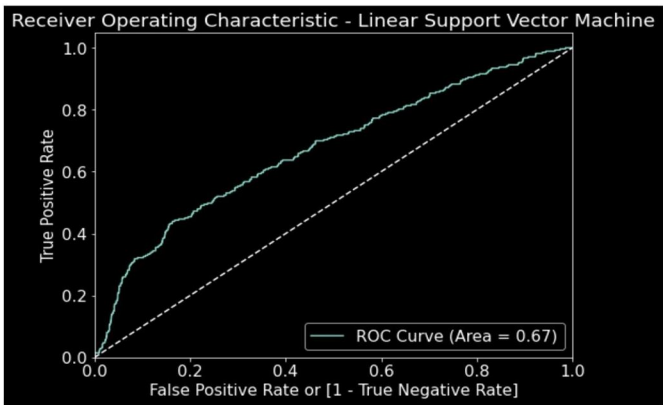


Fig 1. Logistic Regression

F. Random Forest Algorithm

A random forest is an ensemble machine learning algorithm that is used for classification and regression. It is composed of many decision trees, each of which is trained on a randomly selected subset of the data. The predictions made by the individual trees are then combined to make a final prediction, either through a majority vote for classification tasks or through averaging for regression tasks.

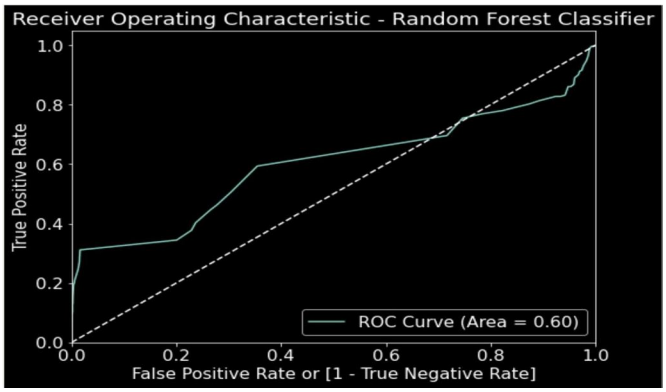


Fig 2. Random Forest Algorithm

G. Gradient Boosting Algorithm

Gradient boosting is a method standing out for its prediction speed and accuracy, particularly with large and complex datasets. From Kaggle competitions to machine learning solutions for business, this algorithm has produced the best results. We already know that errors play a major role in any machine learning algorithm. There are mainly two types of error, bias error, and variance error. Gradient boost algorithm helps us minimize bias error of the model.

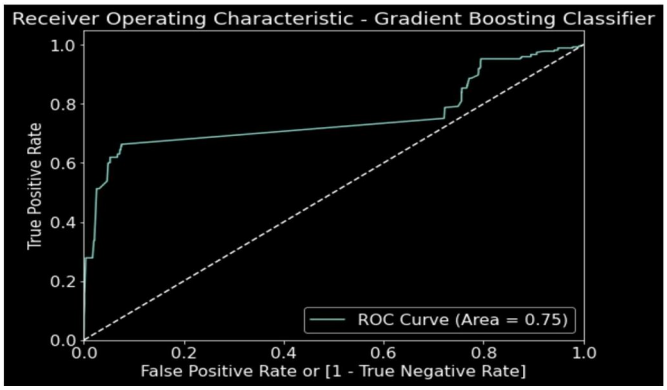


Fig 3. Gradient Boosting Algorithm

H. Predictive Model Result

In this step, we analysis the different algorithms and choosing the best accurate algorithms for the final implementation.

IV. CONCLUSION

At the conclusion of this project, we would say that we learnt a lot from several sources to finish it. We used a variety of techniques to finish this project successfully.

By using machine learning algorithms, precise and accurate identification and detection have become more achievable. Thyroid disease is not easy to diagnosis because mix-up of their symptoms with other condition but our model which is describe in this study give the more effective and more accurate result based on the few inputs.

V. RESULT ANALYSIS

After comparison and analysis of logistic regression, Random Forest Algorithm, Random Forest Algorithm Tuned 1, Tuned 2 and Gradient Boosting Algorithms, it was observed that 89.9% accuracy is achieved by Tuned 1 Random Forest Algorithm in all three parts of experiment, while logistic regression gained second best accuracy 78.92% in Ll feature selection, respectively. Random Forest Algo. also carried out excellent result accuracy of 89.84% with AUC rate of 42.2%.

Upon analyzing the results, the advantages and robustness of new dataset are clearly seen and would allow doctors to get more precise and accurate results in less time.

	Accuracy	AUC
Logistic Regression	0.874378	0.602132
Random Forest	0.894676	0.420194
Random Forest Tuned 1	0.893527	0.591608
Random Forest Tuned 2	0.888548	0.545813
Gradient Boosting	0.411720	0.524825

Fig 4. Result Analysis of algorithms

VI. FUTURE WORK

1) To make the model more practical, different features are also include as like reducing the inputs, applying more algorithms to make model more accurate.
2) A much larger dataset will help in getting a more accurate result.
3) Provide a Wonderful and Interactive Interface to the client for easy use.
4) Develop a new algorithm based on the existing algorithms which give the more accurate result in minimum time with more effectiveness.
5) A mobile application or desktop application could be developed to make the user experience easier.

VII. CODE

We have tried to develop the code on the basis of the requirement of the project so mentioned repository contains the code of the thyroid disease detection system.

https://github.com/honey2438/Thyroid-Disease-Detection

VIII. CONTRIBUTIONS

A. Devansh Sirohi
1) Managing project.
2) Web app development.

B. Deepanshu Kashyap.
1) Preprocessing of auxiliary data and datasets.
2) Training over machine learning techniques.
3) Backend Development.

C. *Devendra Kumar.*
1) Mobile app development.
2) Future work.

D. Gopal Goyal.
1) Methods.
2) Conclusion.

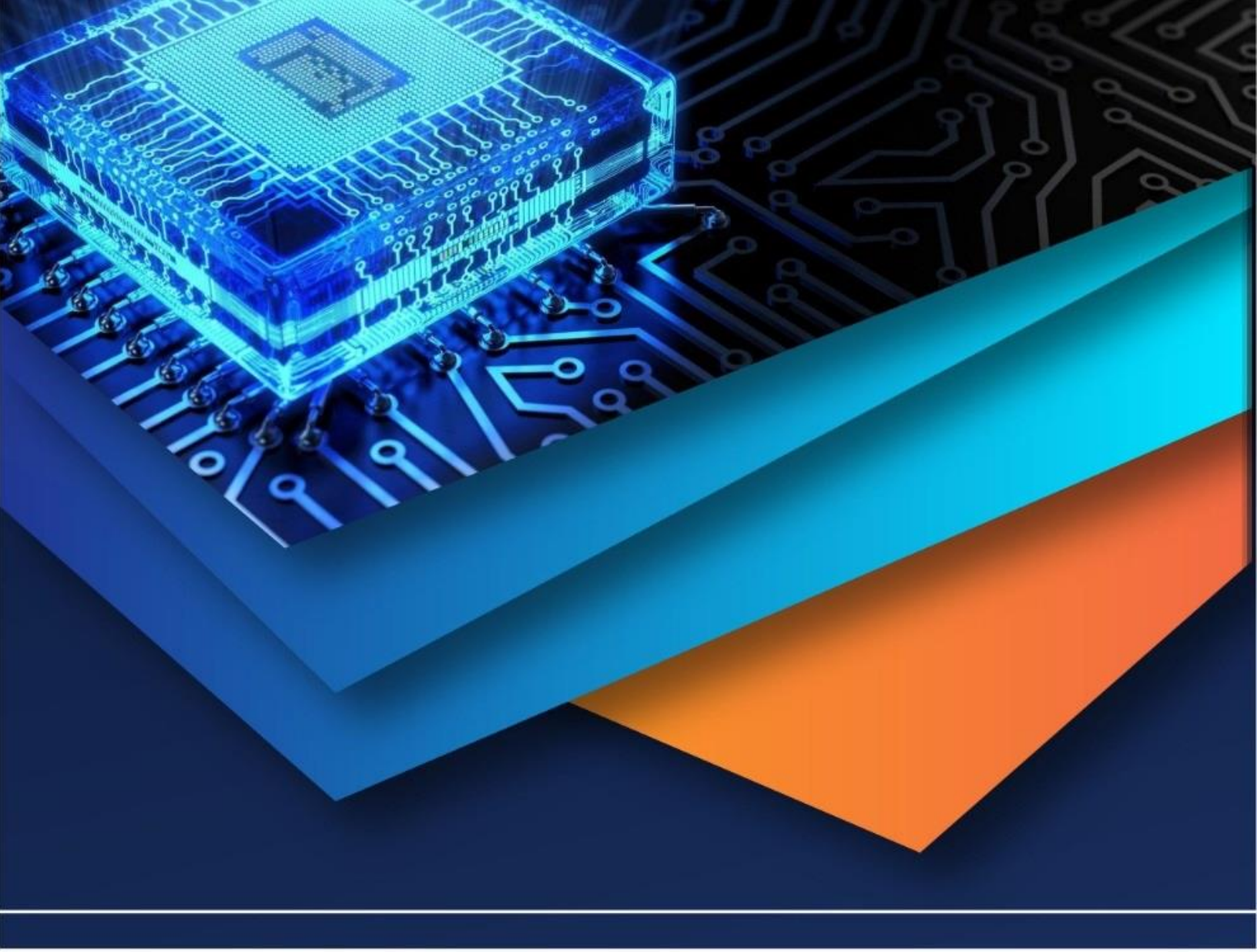
IX. ACKNOWLEDGEMENTS

This is to certify that the research, design, development as well as implementation has been made by all the group members of IMS Engineering College, Ghaziabad.
It's our privilege to express our sincere regards to our project guide, Ass. Prof. Mrs. Bhumica Verma for his valuable inputs, able guidance, encouragement, cooperation and constructive criticism throughout the duration of our project.
We sincerely thank the Project Assessment Committee members for their support and for enabling us to present the project on the topic "Thyroid Disease Detection System".

REFERENCES

[1] Ankita Tyagi and Ritika Mehra. (2018). "Interactive Thyroid Disease Prediction System using Machine Leaming Techniques" published on ResearchGate.
[2] Yong Feng Wang, (2020). "Comparison Study of Radiomics and Deep-Leaming Based Methods for Thyroid Nodules Classification using Ultrasound Images" published on IEEE Access.
[3] Sunila Godara, (2018). "Prediction of Thyroid Disease Using Machine Leaming Techniques" published on IJEE.
[4] Hitesh Garg, (2013). "Segmentation of Thyroid Gland in Ultrasound image using Neural Network" published on IEEE.
[5] Jahantigh, F.F.: Kidney diseases diagnosis by using fuzzy logic. In: 2015 International Conference on Industrial Engineering and Operations Management.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)