DAYANANDA SAGAR UNIVERSITY

KUDLU GATE, BANGALORE – 560068

Bachelor of Technology
in
COMPUTER SCIENCE AND ENGINEERING

Major Project Phase-II Report

DETECTION OF THYROID DISORDER USING MACHINE LEARNING APPROACH

By

R Ajith Kumar (ENG19CS0254)

Under the supervision of
Prof. Meghana G
Assistant Professor, CSE

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING,
SCHOOL OF ENGINEERING
DAYANANDA SAGAR UNIVERSITY,
BANGALORE

(2022-2023)

DAYANANDA SAGAR UNIVERSITY

School of Engineering
Department of Computer Science & Engineering

Kudlu Gate, Bangalore – 560068
Karnataka, India

CERTIFICATE

This is to certify that the Phase-II project work titled **"DETECTION OF THYROID DISORDER USING MACHINE LEARNING APPROACH"** is carried out by **R Ajith Kumar(ENG19CS0254)** bonafide students of Bachelor of Technology in Computer Science and Engineering at the School of Engineering, Dayananda Sagar University, Bangalore in partial fulfillment for the award of degree in Bachelor of Technology in Computer Science and Engineering, during the year **2022-2023**.

Prof Meghana G	**Dr. Girisha G S**	**Dr. Udaya Kumar Reddy K R**
Assistant Professor Dept. of CS&E, School of Engineering Dayananda Sagar University	Chairman CSE School of Engineering Dayananda Sagar University	Dean School of Engineering Dayananda Sagar University
Date:	Date:	Date:

Name of the Examiner	**Signature of Examiner**
1.	
2.	

DECLARATION

I, **R Ajith Kumar (ENG19CS0254)** students of eighth semester B. Tech in **Computer Science and Engineering**, at School of Engineering, **Dayananda Sagar University**, hereby declare that the Major Project Stage-II titled **"Detection of thyroid disorder using Machine learning approach"** has been carried out by us and submitted in partial fulfilment for the award of degree in **Bachelor of Technology in Computer Science and Engineering** during the academic year **2022-2023**.

Student **Signature**

R Ajith Kumar
ENG19CS0254

Place: Bangalore
Date:

ACKNOWLEDGEMENT

It is a great pleasure for us to acknowledge the assistance and support of many individuals who have been responsible for the successful completion of this project work.

First, I take this opportunity to express our sincere gratitude to School of Engineering & Technology, Dayananda Sagar University for providing us with a great opportunity to pursue our Bachelor's degree in this institution.

I would like to thank ***Dr. Udaya Kumar Reddy K R, Dean, School of Engineering & Technology, Dayananda Sagar University*** *for his constant encouragement and expert advice.*

It is a matter of immense pleasure to express our sincere thanks to ***Dr. Girisha G S, Department Chairman****,* ***Computer Science and Engineering****,* ***Dayananda Sagar University,*** *for providing right academic guidance that made our task possible.*

I would like to thank our guide ***Prof. Meghana G****,* ***Assistant Professor****,* ***Dept. of Computer Science and Engineering****,* ***Dayananda Sagar University****, for sparing her valuable time to extend help in every step of our project work, which paved the way for smooth progress and fruitful culmination of the project.*

I would like to thank our ***Project Coordinator Dr. Meenakshi Malhotra*** *and* ***Dr. Pramod Kumar Naik*** *as well as all the staff members of Computer Science and Engineering for their support.*

I am also grateful to our family and friends who provided us with every requirement throughout the course.

I would like to thank one and all who directly or indirectly helped us in the Project work.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ML	Machine Learning
SVM	Support Vector Machine
UI	User Interface

LIST OF FIGURES

ABSTRACT

In India, 42 million people suffer from diseases such as thyroid. Humans have a vascular gland called the thyroid that is one of the most important organs in their bodies. Two hormones are secreted by this gland, which function to control the body's metabolism. When this disorder occurs in the body, certain hormones are released that imbalances the body's metabolism. Using Machine Learning, this project is designed to detect thyroid diseases in humans. Importing the dataset is accomplished through a User Interface. Three different machine learning algorithms are used to construct the model to detect thyroid disease. In this study, SVM, Random Forest, and Naive Bayes are machine learning techniques used to identify thyroid illness. Using three machine learning algorithms, the accuracy of the model is shown. The most effective machine learning algorithm for detecting thyroid disease has been chosen among the various algorithms that have been used.

CHAPTER 1

INTRODUCTION

CHAPTER 1 INTRODUCTION

As we all know, now a days, thyroid is a major and the most frequent disorder mainly among women. The advanced computational biology is widely used in the healthcare industry. This involves gathering and collection of patient details for medical disease detection and prediction. Many intelligent algorithms are used in diagnosing the disease at an early stage. The medical information system is rich with various datasets but the intelligent systems are not available for the easy analysis of the diseases. Ultimately, machine learning algorithms play a major key role in solving the problems with high complexity and non-linear problems while developing a prediction model. The features are selected from various datasets which can be used as the patients details in a healthy patient as accurate as possible that are necessary in any prediction models. Otherwise, misdiagnosis results in an healthy patient that undergoes necessary treatments and care. This thyroid disease is increasing and spreading rapidly all over the world. It is complex to detect this thyroid disorder from the laboratories and requires prior knowledge and good experience.

The thyroid gland is an essential hormone gland which plays a major role in the growth, metabolism and development of the human body. It is a small, butterfly-shaped endocrine gland located in the neck region beneath the Adam's apple which secretes thyroid harmones that effects the metabolism. These thyroid harmones helps in maintaining the body's metabolism as well as the temperature. These harmones also play a role in processing protein and also to burn calories. The types of harmones are T4 (Thyroxine) and T3 (Triiodothyronine) that are released by thyroid gland.

Machine learning plays a major role in detecting this thyroid disease. There are many machine learning algorithms which helps in diagnosing thyroid in a human. This study involves three machine learning algorithms such as Naïve bayes, Support Vector Machine and Random Forest to enhance the prediction accuracy of thyroid, to detect and identify thyroid problems. Thus, if any abnormal conditions of thyroid hormone levels are identified, patients may be prescribed for the treatment and medicine.

1.1. OBJECTIVE

- A large number of data is used to estimate the likelihood of a better result as increasing prediction accuracy will enhance the thyroid problem detection.
- Various pre-processing techniques are applied to enhance the model's performance.
- The accuracy, recall, precision and F1 scores are examined to evaluate the effectiveness of the machine learning algorithms.

1.2. SCOPE

The main purpose of designing the thyroid disease detection system is to create convenient and easy-to-use applications for users to detect thyroid disease. More specifically this system allows the user to self-operate and analyze the thyroid disease contained in them.

1.3. SOCIETAL / ENVIRONMENTAL IMPACT

This study helps doctors accurately predict the likelihood of developing a disease such as thyroid in patients.

CHAPTER 2
PROBLEM DEFINITION

CHAPTER 2 PROBLEM DEFINITION

The symptoms of thyroid disease often vary from person to person and are non-specific, so a proper diagnosis can easily be misdiagnosed.

Finding an accurate solution to this problem for healthcare practitioners via machine learning algorithms for detecting a particular thyroid disease that a person may have will cause an immense decrease in misdiagnoses as it is capable of distinguishing between problems of the thyroid gland.

CHAPTER 3
LITERATURE REVIEW

CHAPTER 3 LITERATURE REVIEW

Over the years, researchers have conducted studies on thyroid diseases. There are numerous research papers that talk about the use of various machine learning algorithms to help diagnose thyroid disease. The research papers that were used as reference for this work have been summarized below.

Hebatullah Mohammad Almahshi, Esraa Abdallah Almasri, Hiam Alquran, Wan Azani Mustafa, Ahmed Alkhayyat proposed a research paper, which includes detection and prediction of hypothyroidism. Machine Learning is used to detect hypothyroid disease. This paper presents an analysis and classification model that takes into account the various factors involved in the prediction of disease. Classifiers used in this paper are support vector machine (SVM), Naive Bayes and decision trees. The results are appeared in the form of three classes (compensated hypothyroidism, primary hypothyroidism, and negative). Future work is considered to increase the number of cases that can be detected and diagnosed using machine learning.

Marissa Lourdes De Ataide and Amita Dessai proposed a research paper, in which thyroid disease is detected using soft computing techniques. Thyroid dataset is collected from UCI repository. Multilayer perceptron Classifier is used for training and classification. Classification of thyroid disease into euthyroid, hyperthyroid and hypothyroid gave an accuracy of 97.5% and further Classification of hypothyroid into primary, secondary and tertiary hypothyroid gave accuracy of 91.7%.

Saima Sharleen Islam, Md. Samiul Haque, M. Saef Ullah Miah, Talha Bin Sarwar and Ramdhan Nugraha - Application of machine learning algorithms to predict the thyroid disease risk: an experimental comparative study. Technology/ design: machine learning algorithms: CatBoost, Extra Trees, ANN, LightGBM, SVC, KNN, Random Forest, XGBoost, Decision Tree, and GaussianNB Results shared by the author Among all the algorithms, the ANN classifier outperforms others with an accuracy of 0.9587. The CatBoost and XGBoost classifiers come second and third with the accuracy of 0.9538 and 0.9533, respectively.

Chandan R, Chethan Vasan, Chethan MS, Devikarani - Thyroid detection using machine learning international journal of engineering applied sciences and technology 2021 classifiers used in this paper are support vector machine(SVM), Decision tree, logistic regression, K-nearest neighbors, Artificial neural network the highest accuracy is for logistic regression algorithm with 90.2%.

Muhammad hamid, Tahir alyas, Khalid Alissa - Empirical method for thyroid disease classification using a machine learning approach. Journal of biomedicine and biotechnology 2001-2012 classifiers used are Decision tree, Random Forest algorithm, KNN and Artificial neural network the highest accuracy is for Random Forest algorithm equal to 94.8% accuracy and 91% specificity.

Lerina Aversaw, Marta Cimitile, Paolo.E. Macchia proposed a study for treatment prediction of thyroid disease where the most used treatments in sodium levothyroxine (LT4), a synthetic thyroid hormone used in the treatment of thyroid disorders. This aims to predict the LT4 treatment for patients suffering from thyroid. In particular, we compared the results of ten different classifiers where among these ten classifiers Extratree classifier showed the accuracy 84%.

Banu, G Rasitha proposed a study where the dataset has been taken at the University of California, Irvine (UCI). They used two data mining techniques J48 and decision stump tree technique. In this paper, the J48 technique was found to be more efficient than the decision stump tree technique. In this analysis, dimensionality reductions are used to pick a subset of attributes from the results. The uncertainty matrix is used to assess classifier output, the J48 algorithm has 96.5% accuracy which is higher than decision stump tree accuracy.

Priyanka Duggal, Shipra Shukla have proposed a paper which presents several methods for classification of thyroid disease diagnosis. The two common diseases of thyroid gland which releases thyroid hormones for regulating rate of body's metabolism are hypothyroidism and hyperthyroidism. To detect and classify the thyroid disorders three classifiers are used i.e., SVM, Naive bayes and Random Forest. Results shows that SVM is the most accurate technique with 92.92% accuracy.

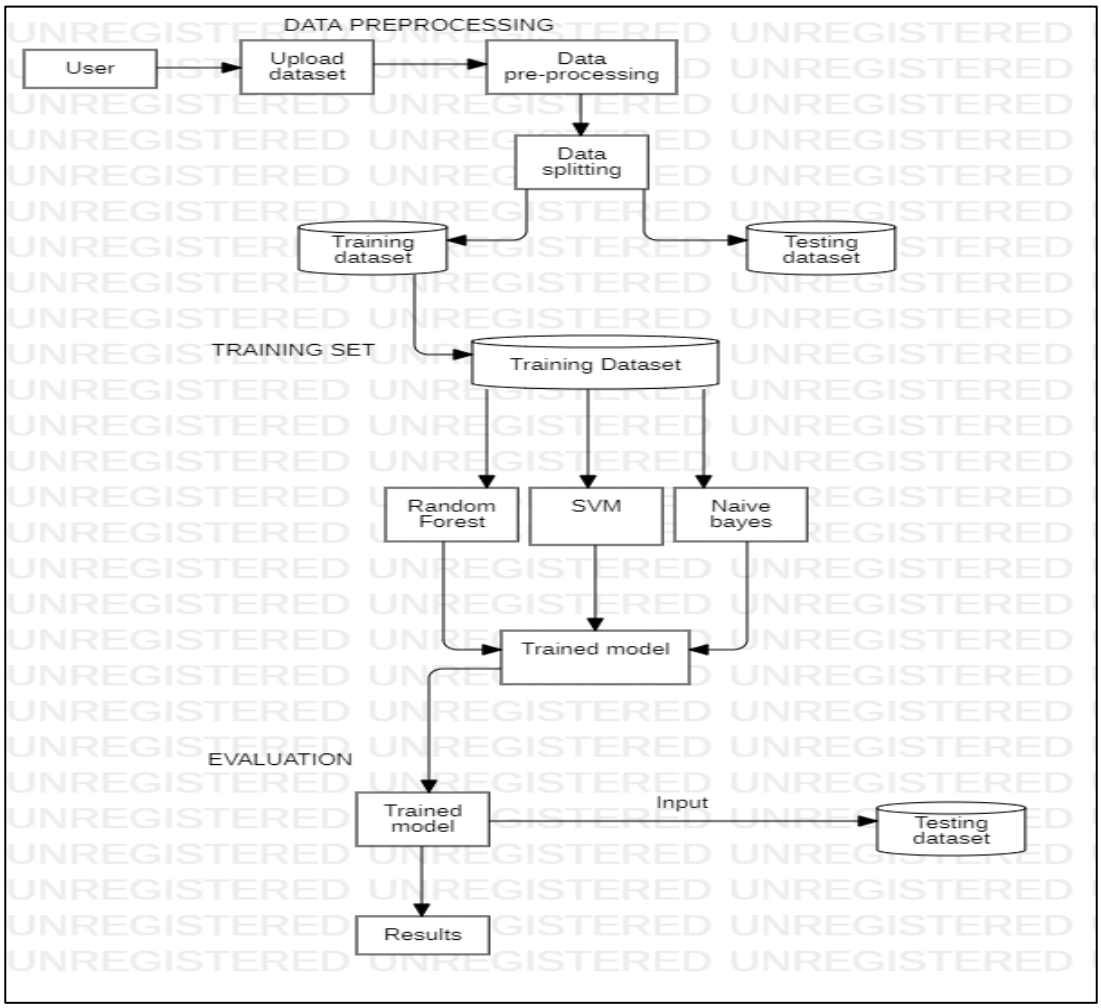
CHAPTER 4
PROJECT DESCRIPTION

CHAPTER 4 PROJECT DESCRIPTION

The main idea of this study is to be able to detect thyroid diseases in human beings. We will be using a dataset which is from Kaggle and then the dataset is used to train our machine learning model. Here in this project, we will be using different machine learning algorithms like Random Forest, Naive Bayes and SVM to get different accuracy as results, it is done so to pick the one with the best accuracy and move on ahead.

4.1. SYSTEM DESIGN

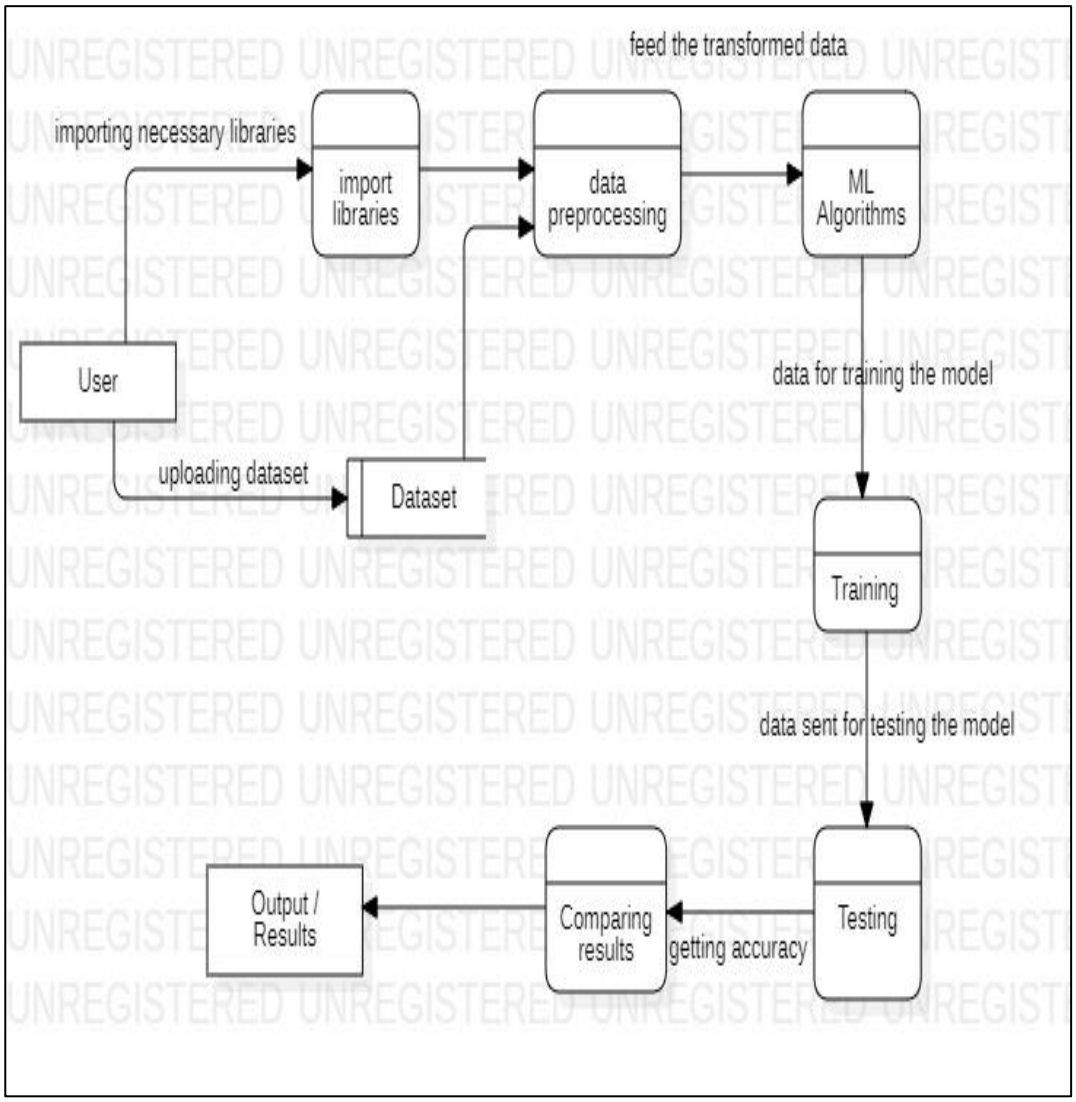
4.1.1. Flow chart diagram



(Fig:4.1.1)

A flowchart is a picture of the separate steps of a process in sequential order. It is a generic tool that can be adapted for variety of purposes and used to describe various processes. In the above flowchart diagram (fig 4.1.1) represents the workflow of the project and also helps to analyze the process of detecting thyroid disease.

4.1.2. Data flow diagram



(Fig:4.1.2)

A data flow diagram is used to represent a flow of data through a process or a system. In the above data flow diagram (fig 4.1.2), represents the flow of the thyroid data and helps in analyzing the detection of thyroid disorder in patients.

CHAPTER 5
REQUIREMENTS

CHAPTER 5 REQUIREMENTS

5.1. FUNCTIONAL REQUIREMENTS

1. The model shall be used to detect thyroid disease from certain parameters which are taken into consideration.
2. This model shall use different machine learning algorithms to detect thyroid disease.
3. Shall depict the accuracy of the model.
4. Shall create a User Interface to implement the model.
5. User shall insert the dataset to get the desired output.

5.2. NON-FUNCTIONAL REQUIREMENTS

1. Availability: Our project should be available at any time of the day.
2. Maintainability: Once the model is trained there shouldn't be any need to maintain unless extra data in the dataset is being added.
3. Reliability: Our project should be accurate enough that the user can rely on the result produced by it.

5.3. HARDWARE & SOFTWARE REQUIREMENTS

Hardware Requirements:

- System: i3 or above
- RAM: 4GB
- Hard Disk: 40GB

Software Requirements:

- Operating System: Windows8 or above
- Coding Language: Python

CHAPTER 6
METHODOLOGY

CHAPTER 6 METHODOLOGY

The initial phase of this project is data collection. It is important to carefully select the data. The data depends on our study aims, objectives, and resource restrictions. The chosen data is next evaluated to prepare it for the model selection procedure. Data pre-processing is done to clear all the unnecessary data from the raw data, which might contain missing, null, and duplicate values. So, to remove all these values, two methods are used. The methods used are LabelEncoder and StandardScaler from Python. After this, the cleaned data is separated into training and testing datasets.

This training and testing data split is used for analyzing the performance of a machine learning system. It is used for problems involving classification and regression and applies to all supervised learning techniques. This process includes separating the dataset into two subgroups. The first subgroup is used to fit the model and is known as the training dataset. In the second subgroup, the input element of the dataset is presented to the model, then predictions are produced and compared to the predicted values. This second subgroup is known as the test dataset. The major goal of this is to evaluate the model's performance on new data. Then, a feature selection procedure is continued. It is the process of limiting the input variables while building a model. It improves the performance of the model and helps to reduce the cost of computing for modeling. It is necessary for the detection and classification of thyroid disorders. Then, the processed data is utilized to implement all the machine learning algorithms Random Forest, Support Vector Machine (SVM), and Naïve Bayes algorithms. Random Forest is a classifier that employs numerous decision trees on various subsets of the input dataset and averages the outcomes to improve the predicted accuracy of the dataset. The SVM algorithm seeks to build the optimum decision boundary or line that can divide n-dimensional space into classes so that we can easily categorize additional data points in the future. Naive Bayes makes predictions based on object probabilities because it is a probabilistic classifier. It uses the Bayes theorem. The formula used is given below.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Here, the Confusion Matrix is used to easily understand the model's performance.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

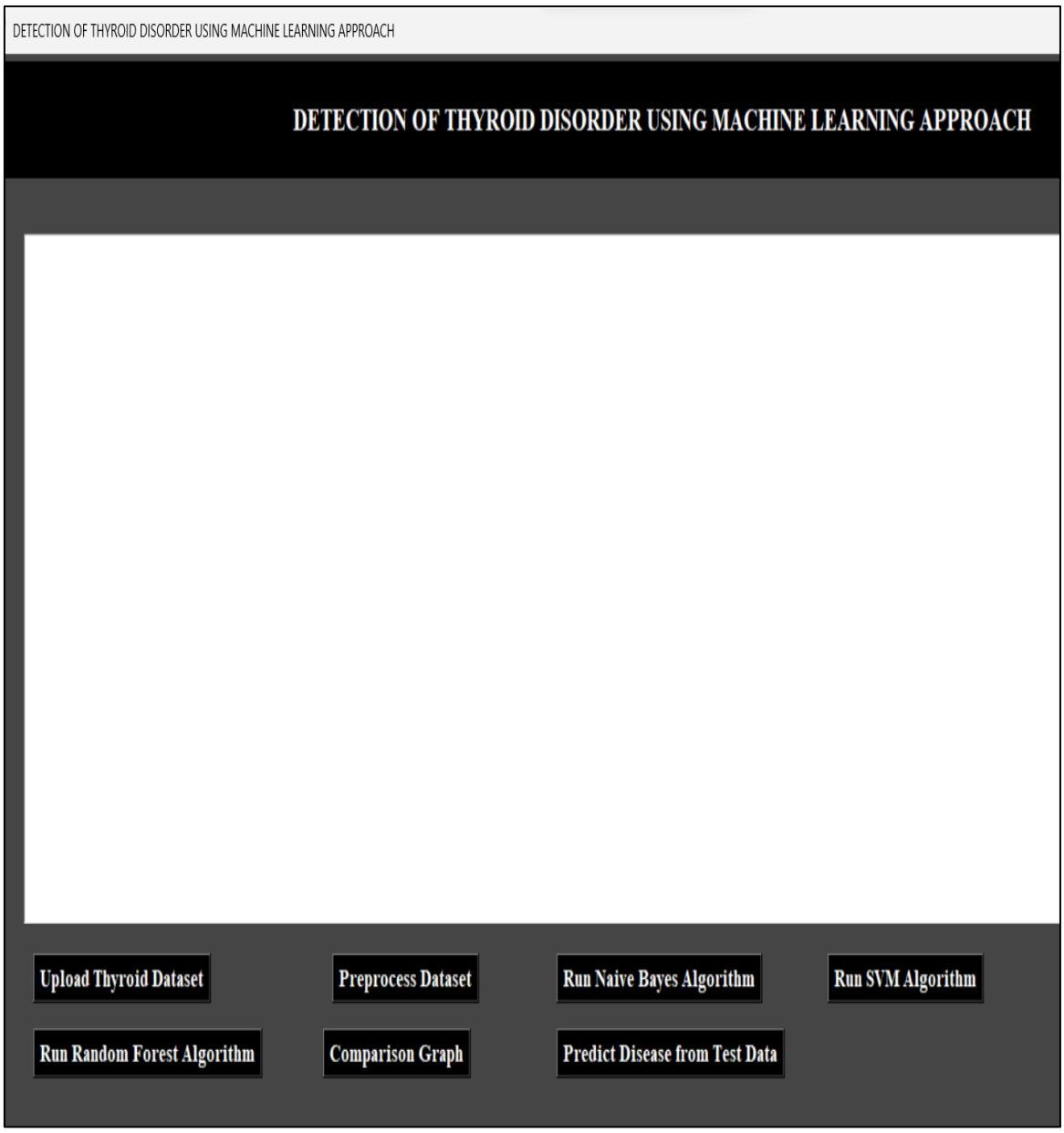
$$F - score = \frac{2 * Recall * Precision}{Recall + Precision}$$

The above formulas are used to find the Accuracy, Precision, Recall, and F-score. The algorithm that provides the highest accuracy is finally selected as the best among the three.

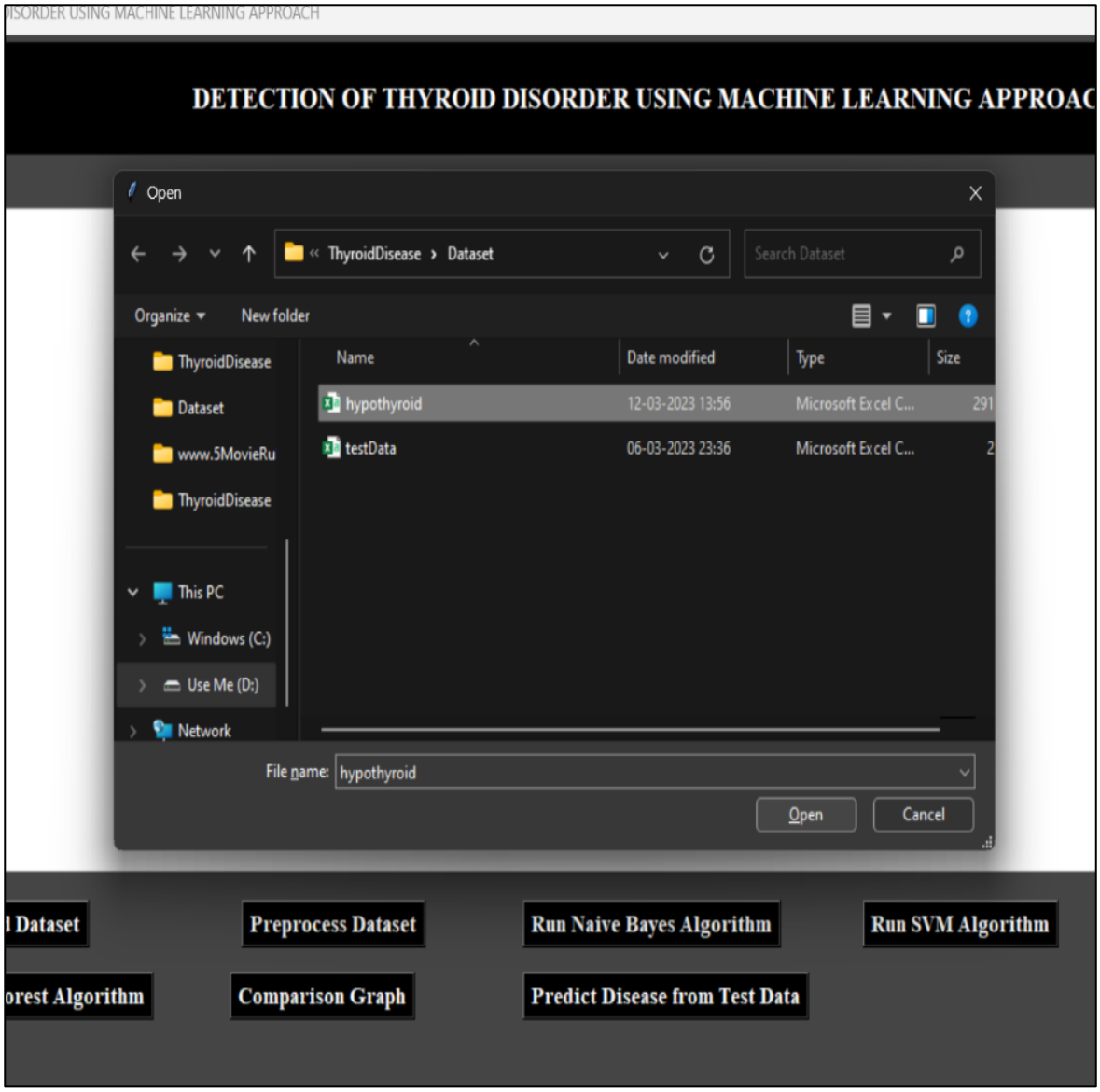
CHAPTER 7
TESTING AND RESULTS

CHAPTER 7 TESTING AND RESULTS

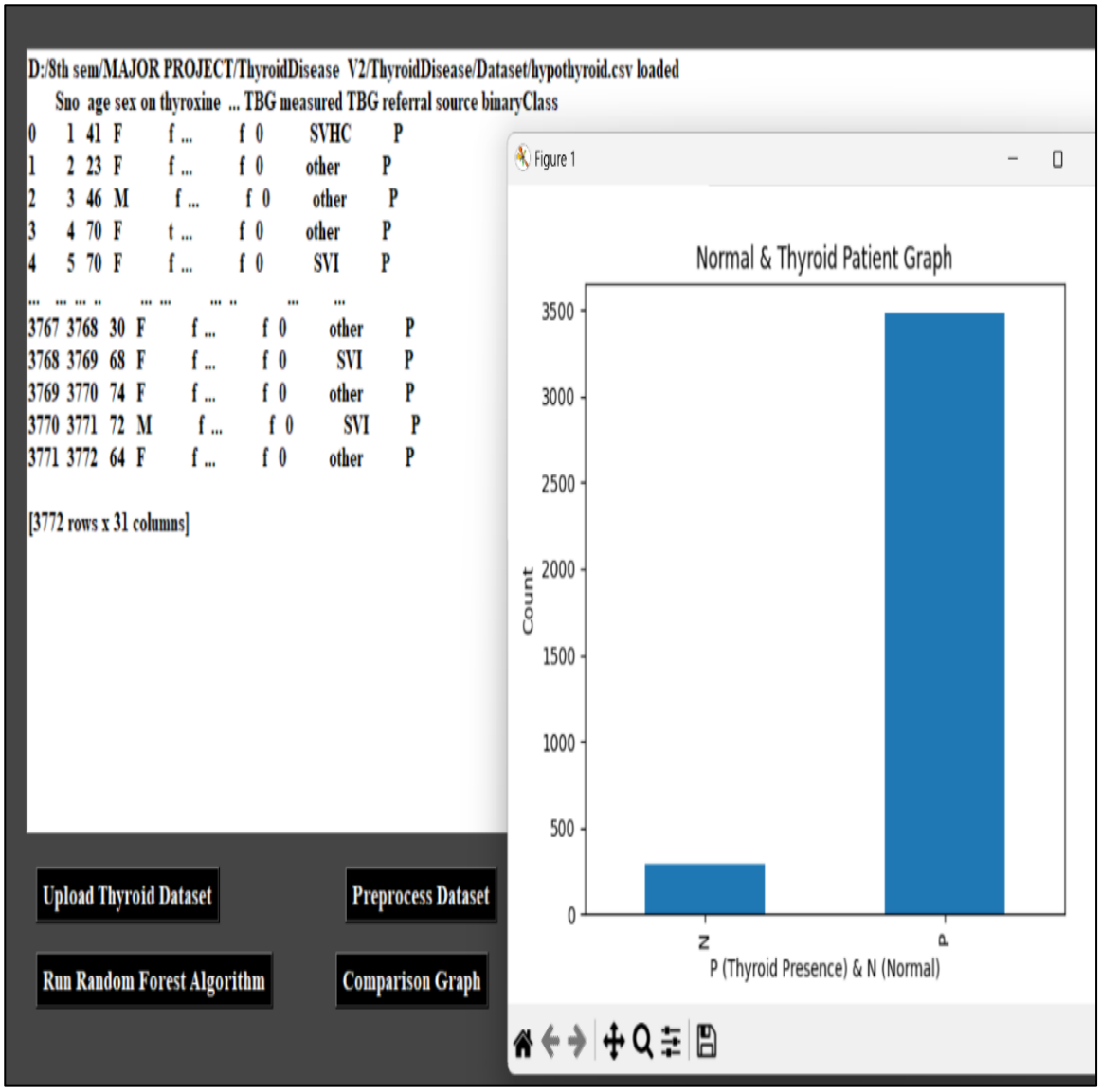
7.1 RESULTS



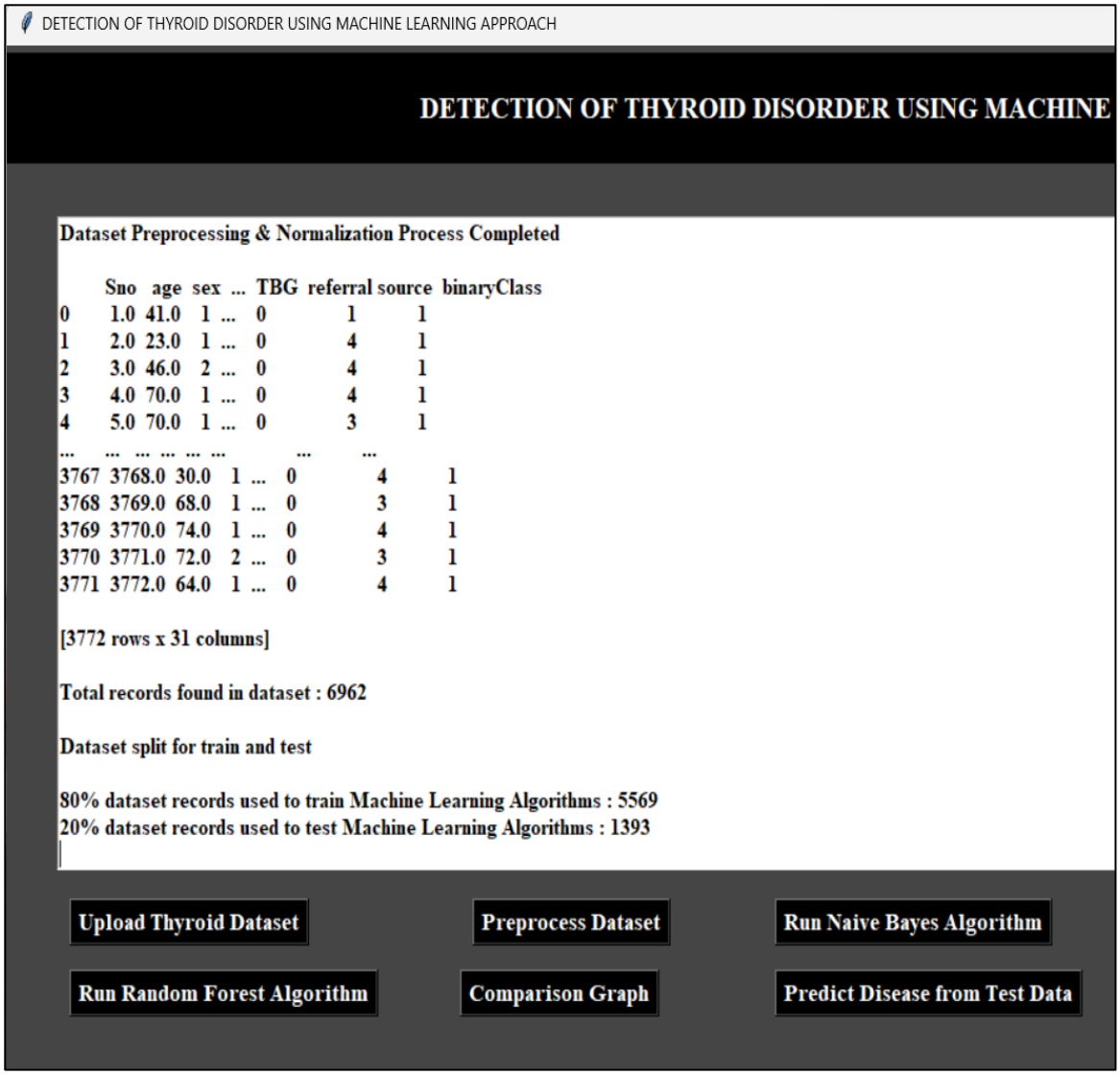
In above screen click on 'Upload Thyroid Dataset' button to load dataset and get below output



In above screen select and upload thyroid dataset and then click on 'Open' button to load dataset and get below output

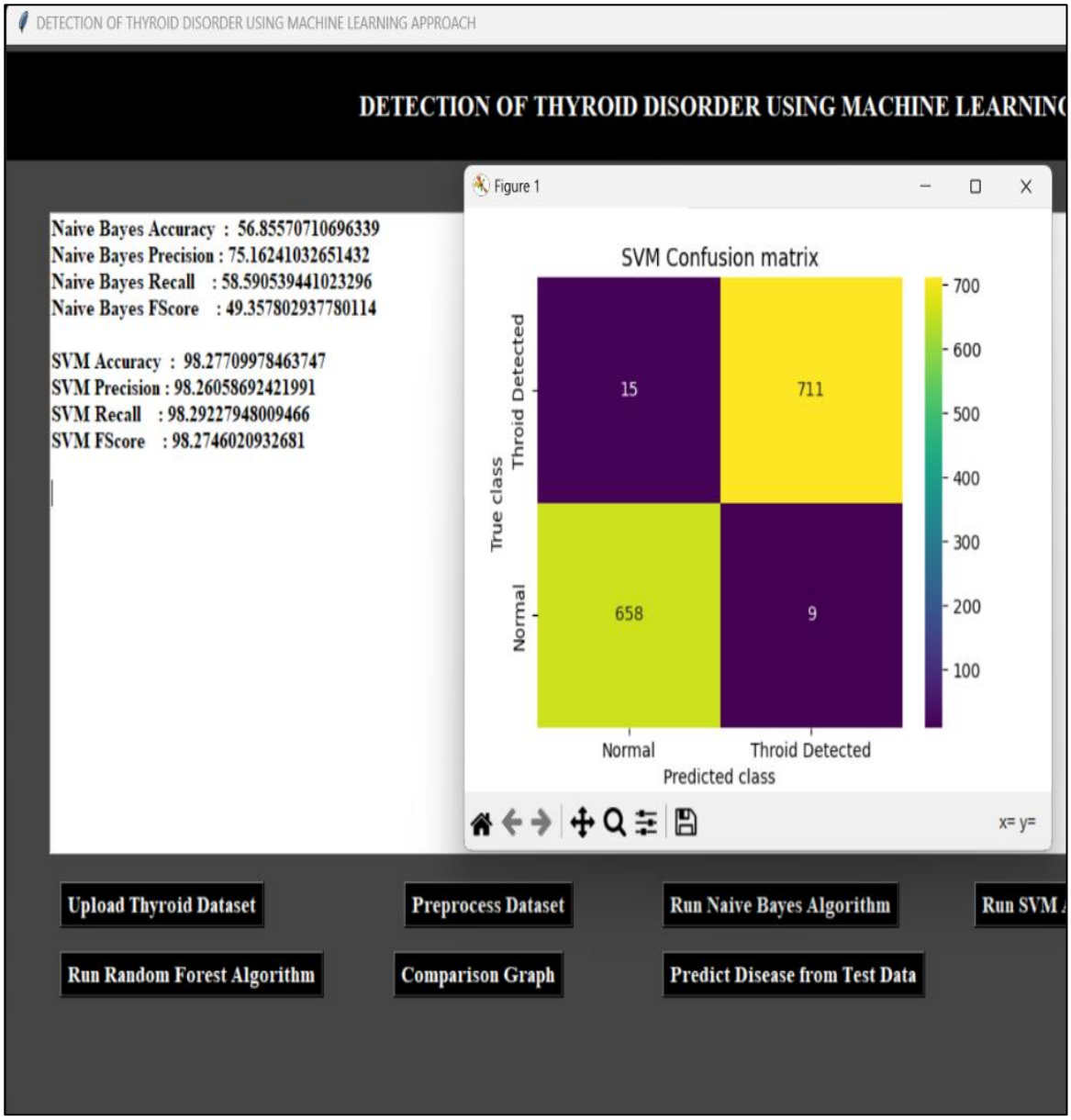


In above screen dataset loaded and in graph x-axis represents N (normal) and P (thyroid presence) and y-axis represents number of records and in above dataset values we can see some are non-numeric and some are numeric and machine learning algorithms accept only numeric values so we need to process dataset to encode non-numeric values to numeric values so click on 'Preprocess Dataset' button to get below output.

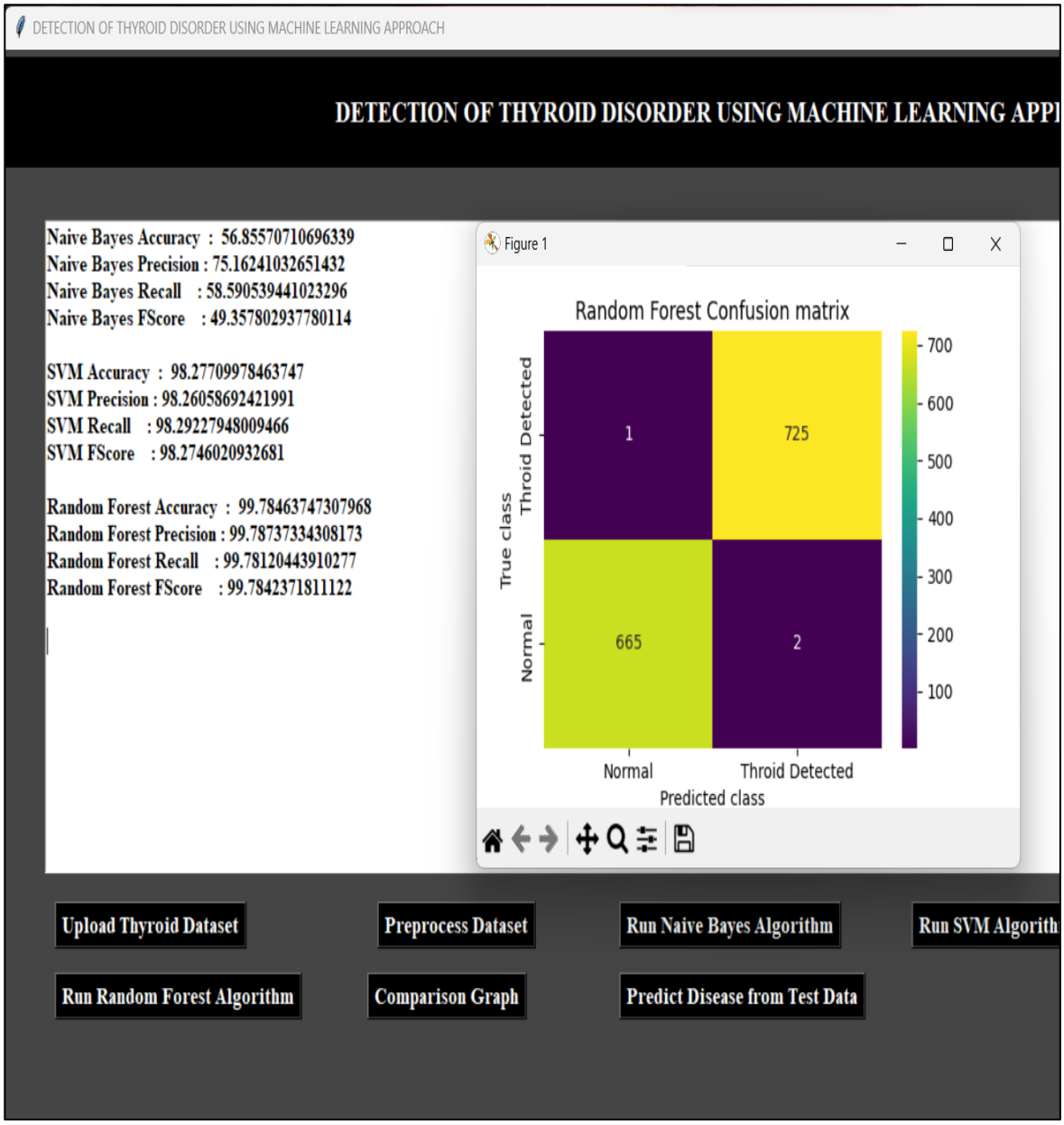


In above screen we can see all values are converted to numeric format and then we can see dataset contains 6962 records where application using 80% records (5569) for training and 1393 (20%) records for testing. Now click on 'Run Naïve Bayes Algorithm' button to train Naïve Bayes on 80% dataset and test on 20% data to get below prediction accuracy

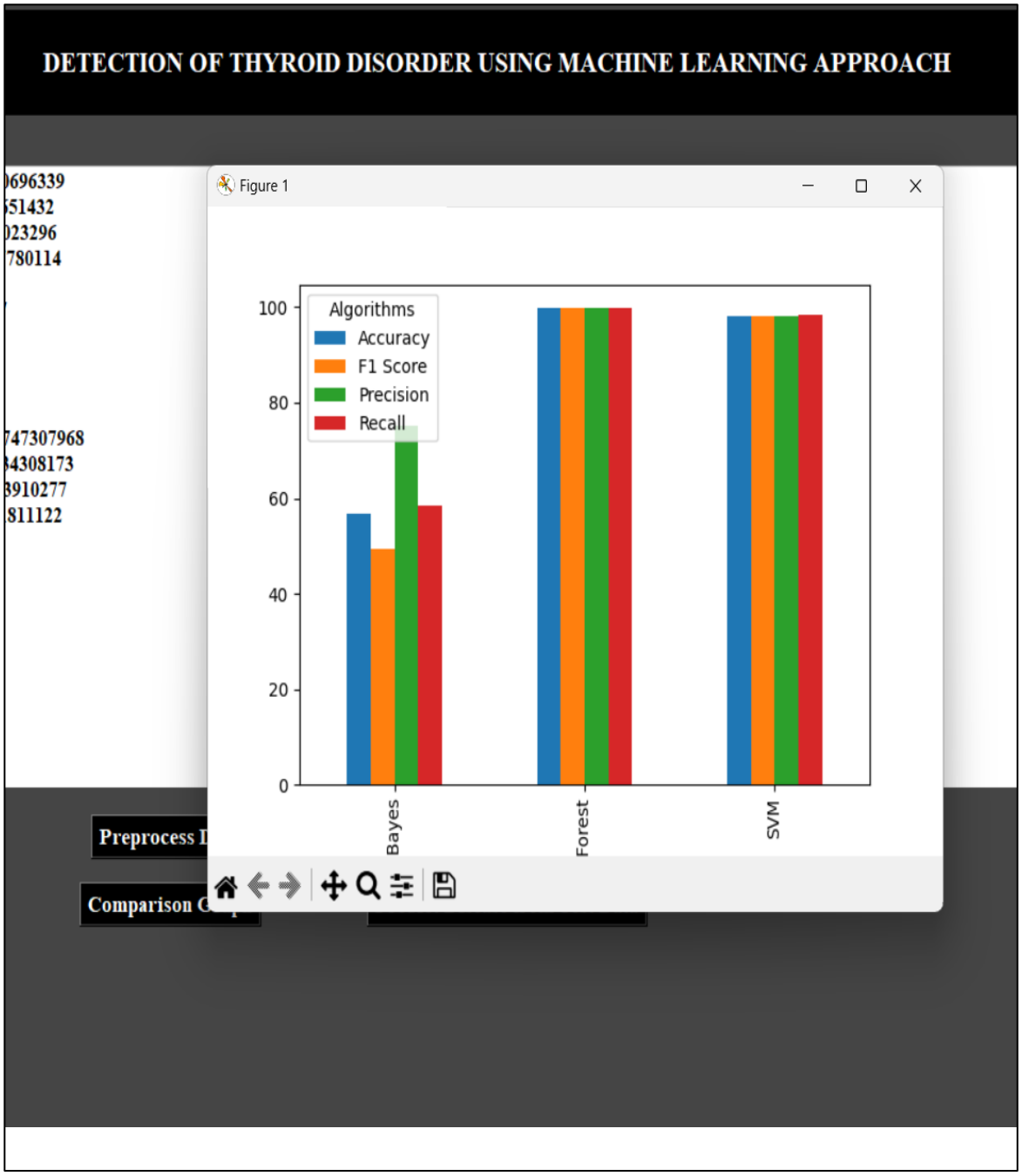
In above screen with Naïve Bayes, we got 56.8% accuracy and we can see other metrics like precision, recall and F-score. In confusion matric graph x-axis represents Predicted Labels and y-axis represents True Labels and yellow and light blue color in diagonal represents correct prediction and dark blue and green box contains incorrect prediction count. Now close above graph and then click on 'Run SVM Algorithm' button to get below output.



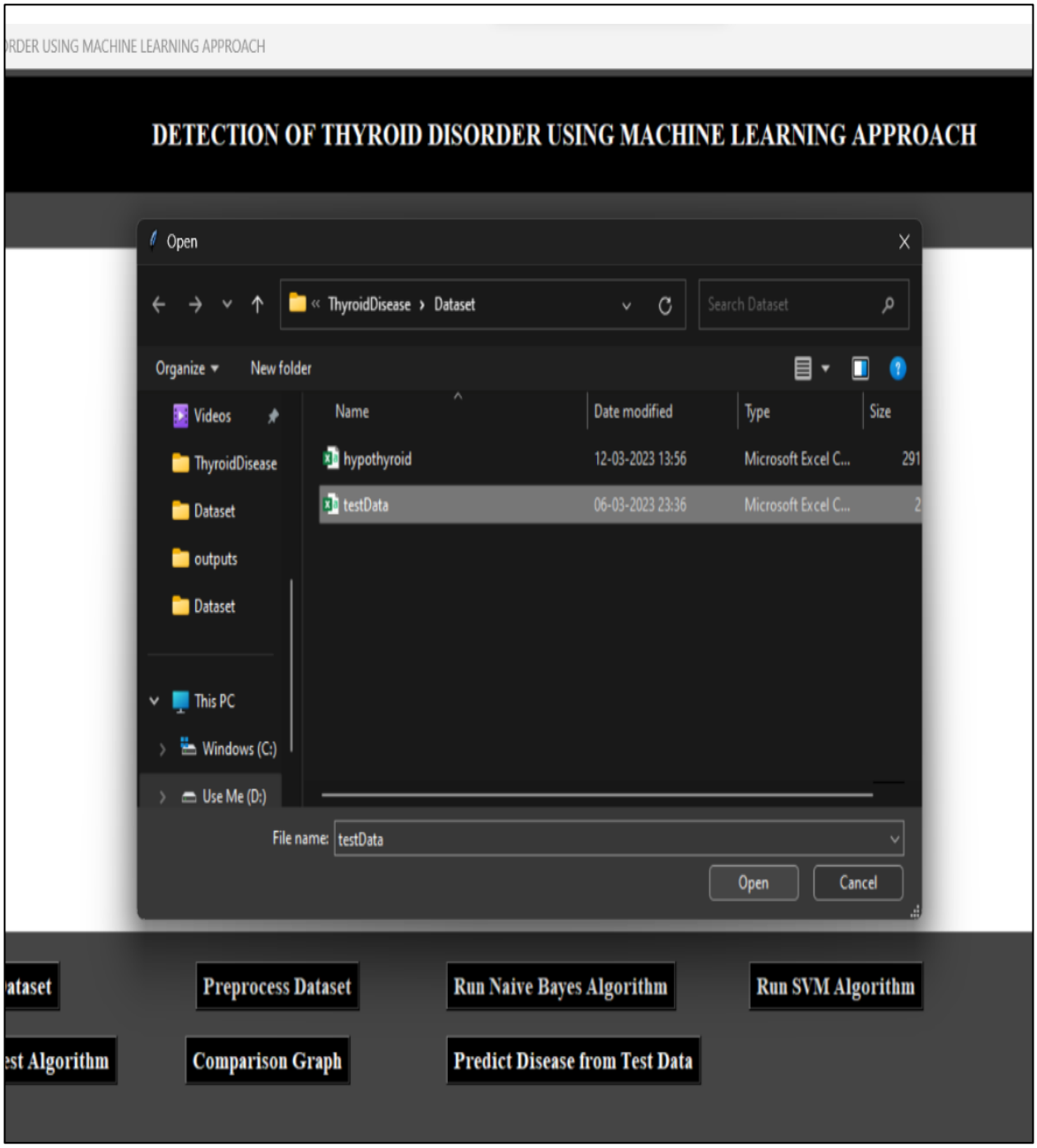
In above screen with SVM we got 98.2% accuracy and in confusion matrix graph yellow boxes contains correct prediction count and blue boxes contains incorrect prediction count. Now close above graph and then click on 'Run Random Forest Algorithm' button to get below output



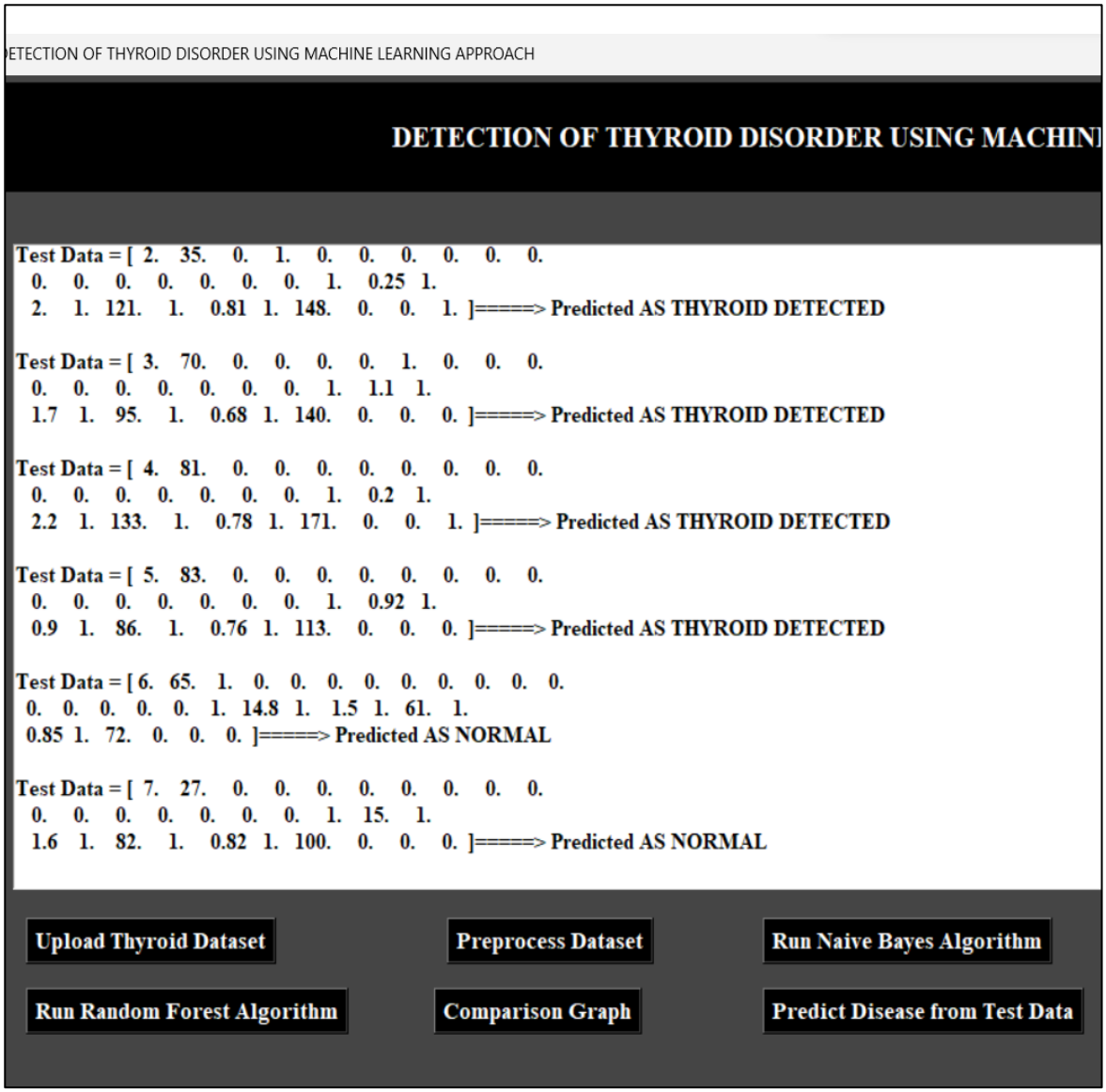
In above screen with Random Forest, we got 99.7% accuracy and now click on 'Comparison Graph' button to get below graph



In above graph x-axis represents algorithm names and y-axis represents metric values like accuracy, precision, recall in different bar color and in all algorithms Random Forest got high accuracy. Now close above graph and then click on 'Predict Disease from Test Data' to upload test data and get prediction output.



In above screen selecting and uploading testData.csv file and then click on ‘Open’ button to get below output.



In above screen in square bracket, we can see test data and after arrow symbol we can see predicted output. Thyroid disease is detected using Random Forest, SVM and Naïve bayes algorithms. Accuracy of three machine learning algorithms used is depicted. Random forest gives the highest accuracy among other machine learning algorithms with 99.7%.

CHAPTER 8
CONCLUSION

CHAPTER 8 CONCLUSION

The study for the Naive Bayes algorithm, Random Forest, and Support Vector Machine to predict thyroid illnesses is presented in this publication. In order to forecast thyroid illnesses, these algorithms are applied to a dataset. Precision, recall, F1 score, and accuracy are calculated to evaluate the algorithms used. The Random Forest algorithm performs best overall, with an accuracy rate of 99.8. With accuracy of 98.7, the Support Vector Machine comes in second. Finally comes the Naive Bayes method, which has the lowest accuracy at 57.9. With the help of this result while predicting thyroid diseases Random Forest is used for accurate result. We also draw the conclusion that medical personnel can use the suggested model as a tool to more precisely diagnose patients. This could aid in avoiding the physical labor that might produce wrong findings and be quite time consuming.

CHAPTER 9
REFERENCES

CHAPTER 9 REFERENCES

[1] Lerina Aversanoa, Mario Luca Bernardia, Marta Cimitileb, Martina Iammarinoa, Paolo Emidio Macchiac, Immacolata Cristina Nettorec, Chiara Verdonea."Thyroid Disease Treatment prediction with machine learning approaches". 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, Procedia Computer Science 192 (2021) 1031–1040.

[2] Priyanka Duggal, Shipra Shukla. " Prediction Of Thyroid Disorders Using Advanced Machine Learning Techniques ". 10th International Conference on cloud computing, data science and engineering (Confluence).

[3] Marissa Lourdes De Ataide, Amita Dessai. "Thyroid Disease Detection using Soft Computing Techniques". International Research Journal of Engineering and Technology (IRJET)

[4] Hebatullah Mohammad Almahshi, Esraa Abdallah Almasri, Hiam Alquran, Wan Azani Mustafa, Ahmed Alkhayyat. "Hypothyroidism Prediction and Detection Using Machine Learning". 5th International Conference on Engineering Technology and its Applications 2022- (5thIICETA2022) 978-1-6654-7215-9/22/$31.00 ©2022 IEEE.

[5] Dr. G. Rasitha Banu, M.Baviya, Dr.Murtaza Ali3. "A STUDY ON THYROID DISEASE USING DATA MINING ALGORITHM". International Journal of Technical Research and Application e-ISSN: 2320-8163, www.ijtra.com volume 3, Issue4 (July August 2015), PP. 376-379

[6] Ankita Tyagi, Ritika Mehra, Aditya Saxena. "Interactive Thyroid Disease Prediction System Using Machine Learning Technique ".5th IEEE International Conference on Parallel, Distributed and Grid Computing (PDGC-2018), 20-22 Dec 2018, Solan, India.

[7] Md Riajuliislam, Khandakar Zahidur Rahim, Antara Mahmud."Prediction of Thyroid Disease (Hypothyroid) In Early Stage Using Feature Selection and Classification Techniques". 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), 27-28 February, Dhaka.

[8] Shaik Razia1* and M. R. Narasinga Rao2." Machine Learning Techniques for Thyroid Disease Diagnosis – A Review". Indian Journal of Science and

Technology, Vol 9(28), DOI: 10.17485/ijst/2016/v9i28/93705, July 2016.

[9] Tahir Alyas, 1 Muhammad Hamid,2 Khalid Alissa,3 Tauqeer Faiz,4 Nadia Tabassum, And Aqeel Ahmad6 "Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach "Hindawi BioMed Research International Volume 2022, Article ID 9809932, 10 pages https://doi.org/10.1155/2022/9809932.

[10] Devansh Sirohi1, Deepanshu Kashyap2, Devendra Pal3, Gopal Goyal4, Bhumica Verma IMSEC Ghaziabad."Thyroid Disease Detection System", International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538 www.ijraset.comXXX.- Volume 11 Issue I Jan 2023.