# THYROID DISEASE DETECTION USING SVM ALGORITHMS

[1]Ashwini Varma, [2]Y. Susheela

[12]Vaageswari College of Engineering, Karimnagar, Telangana, India

[1]ashwinivarma1100@gmail.com [2]chiduralasusheela@gmail.com

**ABSTRACT:** Diseases of the thyroid gland are a crucial factor in medical diagnosis and prognosis, which is a difficult concept in the medical profession. One of the most critical parts of the human body is the thyroid gland. Thyroid hormones have a key role in metabolic regulation. Thyroid hormones play a role in the body's capacity to control its metabolism, and both excess and deficiency can have negative effects. Predicting illness and researching thyroid disease categorization models using hospital datasets are both vital applications of machine learning. A good knowledge base, built and applied as a hybrid model, is essential for dealing with dynamic learning tasks like medical diagnosis and prediction. Using basic machine learning techniques, it may be possible to identify and suppress thyroid activity. The use of a support vector machine (SVM) model for predicting the likelihood of a thyroid patient is commonplace. In cases where a patient is at danger for developing thyroid disease, our system must provide recommendations such as home remedies, warnings, precautions, prescriptions, etc.

## 1. INTRODUCTION

Some of the most cutting-edge applications of machine biology are in medical care. Gathering information for medical complaint vaticination was essential. Colorful intelligent vaticination algorithms are used to find problems with a product or service early on. Although the Medical Information System excels at handling data sets, there are currently no intelligent technologies available for providing a prompt prognosis of patients' ailments. In the end, machine literacy algorithms play a significant role in tackling difficult and non-linear challenges throughout the development of the vaccination model. Any complaint vaticination model must require naming qualities from colourful data sets that may be employed as description in a healthy situation as precisely as feasible. A misclassification may lead to a happy case

being placed in a bad care setting. It's also of the utmost cardinal importance that any possibility of vaccinating against thyroid disease be considered. The thyroid gland is a gastrointestinal endocrine gland. It's built in the human neck's lower region, under the Adam's apple, and helps the body store thyroid hormones, which in turn impacts the body's basal metabolic rate and its ability to produce proteins. These hormones rely on the rate at which the heart beats and the calories are burned to regulate the body's metabolism. Thyroid hormones play a role in regulating the body's metabolism through their chemical makeup. These organs produce the hormones thyroxine (shortened to T4) and triiodothyronine, which are responsible for regulating the body's metabolism ( shortened T3). Thyroid hormones have a crucial role in industrial processes, as well as in the building and maintenance of structures and the regulation of body temperature. Two activated thyroid hormones, T4 and T3, make up the thyroid glands. These hormones have a crucial role in regulating proteins, maintaining internal body temperature, transporting energy throughout the body, and promoting cell proliferation. Iodine, along with T3 and T4 harmones, is a fundamental building block of the thyroid glands and is only inactive in a handful of extremely rare yet urgent situations. Both hypothyroidism and hyperthyroidism can result from insufficient levels of these hormones in the body. There is a wide range of possible causes for both overactive and underactive thyroid. Several pharmaceutical options exist. Iodination deficiency, thyroid atrophy, and a lack of thyroid hormone-producing enzyme are all side effects of thyroid surgery.

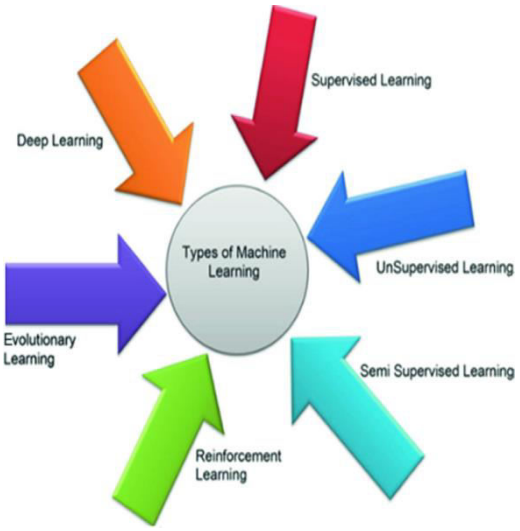



Fig.1: Machine learning techniques for Thyroid detection

## 2. LITERATURE REVIEW

A natural solution to this problem would be to employ machine learning-based techniques to automatically detect and categorise malicious software from unidentified binary code through static analysis. This research follows the

guidelines and makes use of associated technologies of machine learning based approaches to investigate the categorization of malware using this approach [11–14]. Malware detection is fundamentally a classification problem that seeks to label samples as either malicious or benign. This paper's primary contributions to the field of study that underpins the host malware detection technology are as follows: 1 Acquire a sizable collection of both malicious code and genuine software to analyse. Next, process the data and pull out the features from the sample efficiently. Third, zero focus on the most crucial characteristics for a classification. Fourth, construct a classification model by combining the training with machine learning methods. 5. Make use of the learned classification model to spot unidentified samples. Learning which features and models are most useful for this specific application is the end aim. The fundamental concepts and key research issues are presented in this chapter. To begin, let's define a few terms that will be introduced below.

**2.1X. Gao, J. Qiu*, Z. Tian*, S. Su, and Y. Sun. An IoT Anomaly Detection Feature Selection Strategy Based on Correlation Shifts. The Real World Science.**

Sensors are increasingly being installed on machines and their operational condition is being analysed in this age of the fourth industrial revolution. The rapid evolution of IoT technology [1] has made it possible to readily retrieve sensor data from industrial equipment, evaluate it at the network's edge or in data centres, and then use that insight to improve industrial control. As a result of deep learning's rapid advancement in recent years, it is now routinely used for this type of analysis [2, 3, 4]. These techniques pick out a portion of the fetched sensor data stream as the input features, and then anticipate how pieces of machinery will perform. Features chosen had a significant effect on the learning model's performance; hence, feature selection is essential for these kinds of approaches.

Researchers attempt to pick the most relevant features to the prediction model to enhance the prediction performance, or the most informative characteristics to undertake data reduction, while deciding on a collection of features for a learning model. While both types of approaches have their uses, they each have flaws when applied to online contexts. Specified evaluation criteria, such as feature relevance metrics [5] or a predefined learning model [6], are crucial for the former kind of approaches. As a result,

such approaches can only be used with a limited collection of data and are inappropriate for dynamic, unsupervised feature selection in real-time online applications. The latter approaches work wonderfully in digital settings. However, the most important aspect of live industrial equipment status analysis is not efficiency (but rather accuracy), and this is where data reduction comes in.

Researchers are shifting away from relying on predefined evaluation criteria in favour of selecting features that can represent the characteristics of the online sensor data, such as features with the highest cluster ability [8,9] or the smoothest features on the graph [7]. In this study, we examine the characteristics of classic pattern recognition areas, such as image processing and speech recognition [7,8,9], with a particular emphasis on the features with correlation changes, such as smoothness and cluster ability. We think that shifts in correlations can help to accurately identify developments in the industrial environment. This is the first published work investigating the role of correlation shifts in online feature selection, to the best of our knowledge.

**2.2.Thanks to X. Yu, Z. Tian, J. Qiu, and F. Jiang. Reduced Confidential and Contextual Terms as a Data Leakage Prevention Strategy for Smart Mobile Devices, Wireless Communications and Mobile Computing, https://doi.org/10.1155/2018/5823439.**

Information leakage on smart mobile devices is a growing concern [1, 2] as the Internet and information technology permeate more aspects of daily life and bring about the widespread adoption of "smart" mobile devices. Intellectual property and financial data are only two examples of the kind of sensitive information that could fall into the wrong hands, but it could happen with any type of leak. Once sensitive information has leaked, it cannot be stopped from being shared.

Most data breaches, according to surveys [3, 4], originate from within an organisation. Around 29% of these attacks originate from within the company, most often in the form of inadvertent leaks of private or sensitive data; 16% come from theft of intellectual property; and 15% come from other thefts, such as customer information and financial data. In addition, over 67% of businesses agree that internal dangers are more dangerous than external ones.

It is difficult to prevent data leakage efficiently, despite the fact that laws and regulations have been created to punish various acts of deliberate data leakage.

Rephrasing secret contents or embedding confidential contents in non-confidential contents are simple ways to hide sensitive information [5, 6]. Numerous software and hardware solutions have been created to address the issues caused by data leaking, and these will be explored in the following chapter.

In this research, we introduce CBDLP, a model for preventing data leaks that makes use of both confidential phrases and the terms that provide context for them in order to effectively identify rephrased sensitive content. As part of CBDLP, documents of the same class are represented using a graph structure that incorporates confidential terms and their context, and a document's confidentiality score is then calculated to provide evidence for the presence or absence of confidential content. We also suggest a pruning strategy based on the attribute reduction technique used in rough set theory. After pruning, the graph structure of each cluster is modified based on how pivotal the privileged phrases are and the context in which they are used. The purpose of this study is to provide a method that can successfully stop insider data leaking, whether it be purposeful or inadvertent. Mixed-confidentiality documents are widespread, thus it's crucial to precisely identify those that include confidential information, even if it's been rephrased.

## 3. IMPLEMENTATION

Data from the UCI machine learning archives is used in both the research papers and the model classifications used in the prediction of thyroid disease. A good knowledge base that can be centralised and used as a hybrid paradigm must be maintained in order to address complex learning concerns like medical diagnostics and statistical tasks. Additionally, we provided a selection of machine learning and thyroid diagnostic approaches. Thyroid disease risk was estimated using machine learning algorithms like the Vector Support Machine.

In this project, we are using a support vector machine learning algorithm called SVM to determine if a patient's reported data is normal or indicative of a higher risk for thyroid disease; if the latter is the case, the application will present the patient with appropriate dietary and prescription guidelines. In this study, we are developing a prediction model by training a support vector machine using the UCI machine learning THYROID illness dataset. The purpose of applying new patient test data to a trained SVM model is to predict whether a patient is normal or at risk of thyroid disease.

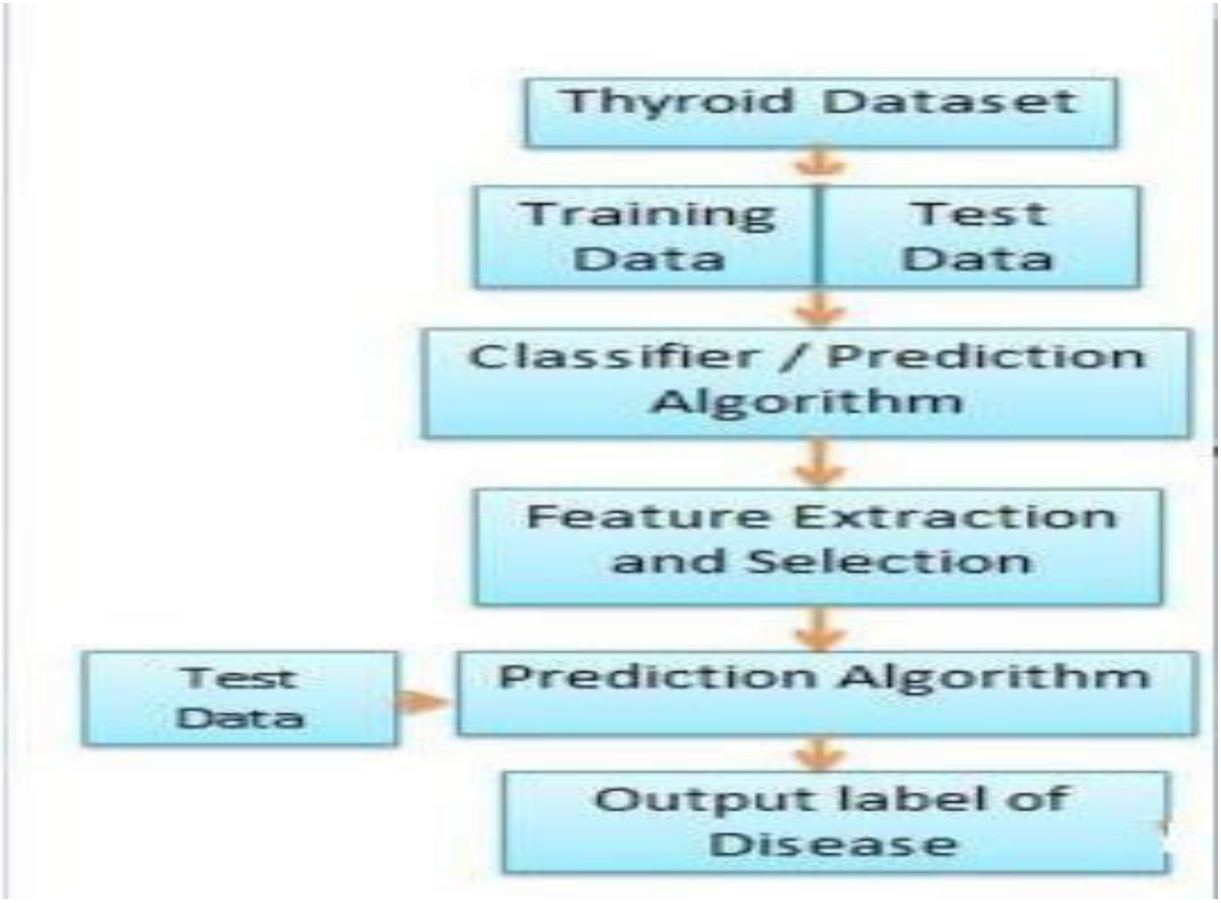



**Fig.2: Workflow diagram**

## 4.DATASET

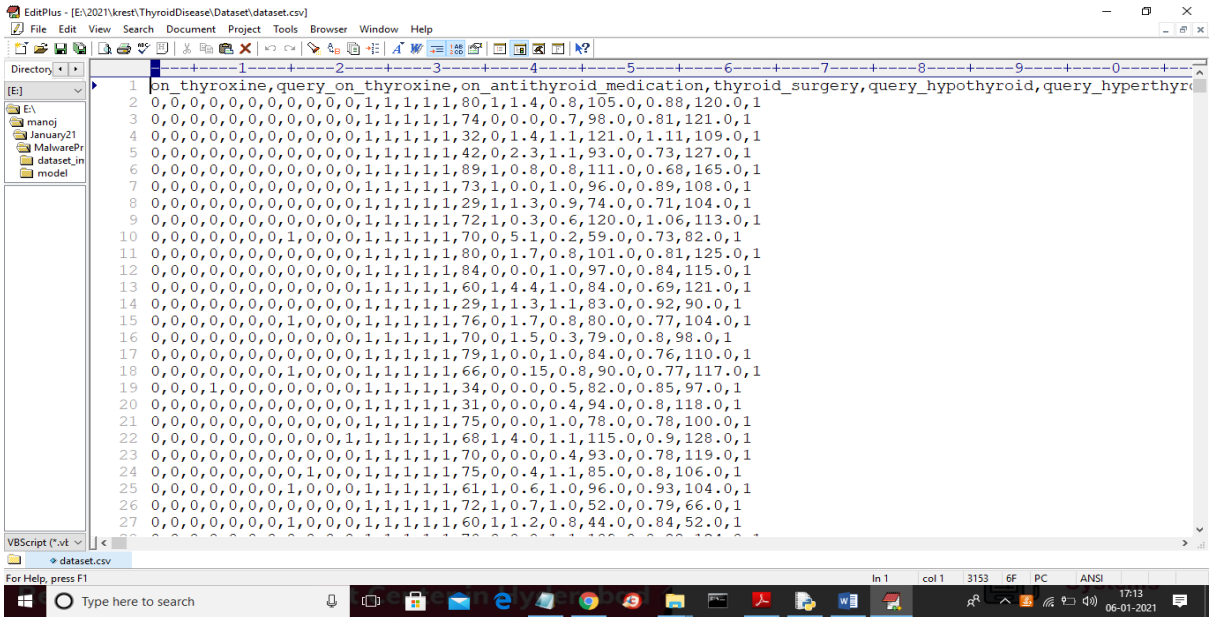
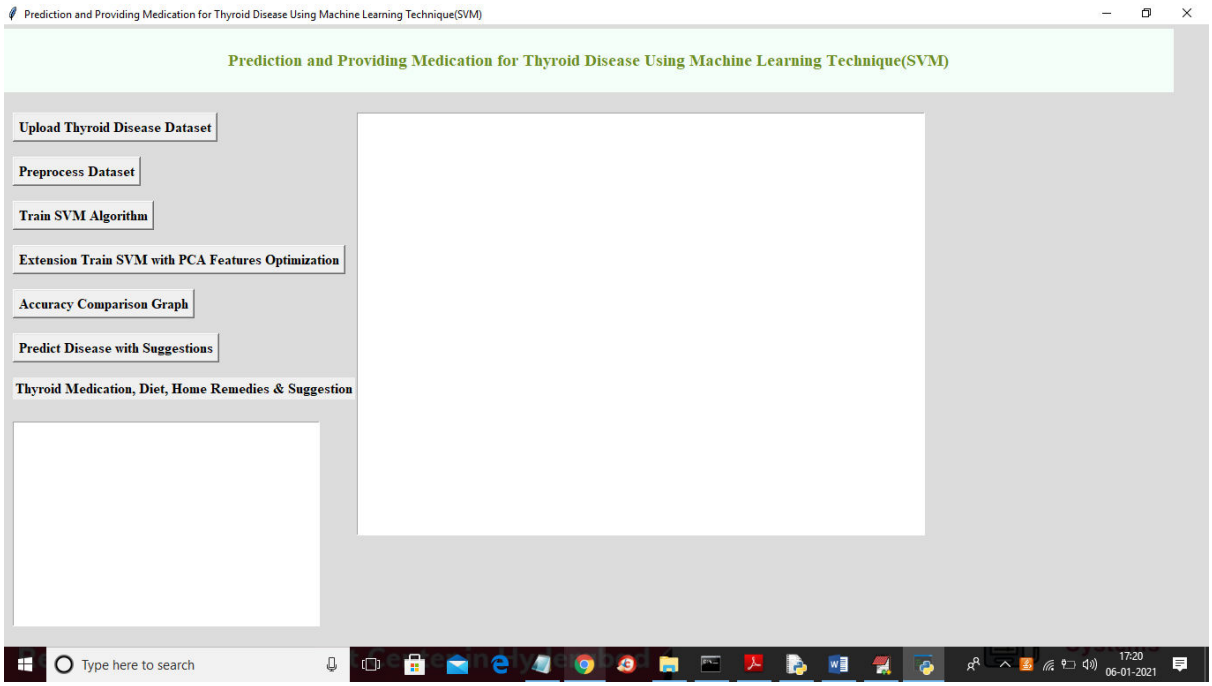
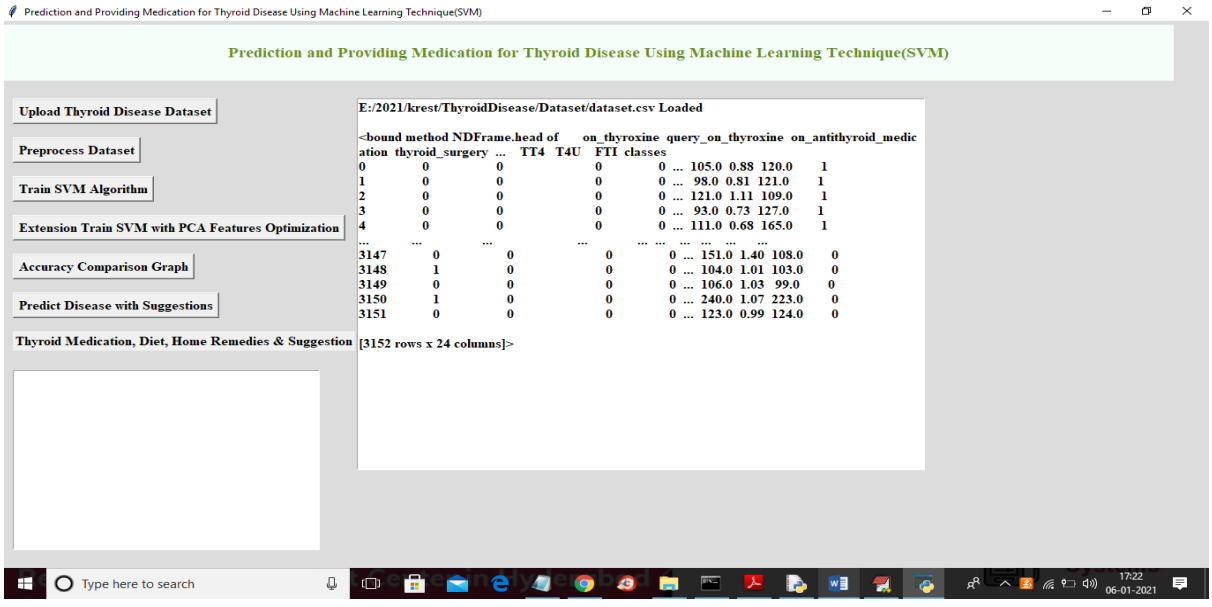


**Fig 4:Data Set Values**

## 5. EXPERIMENTAL RESULTS

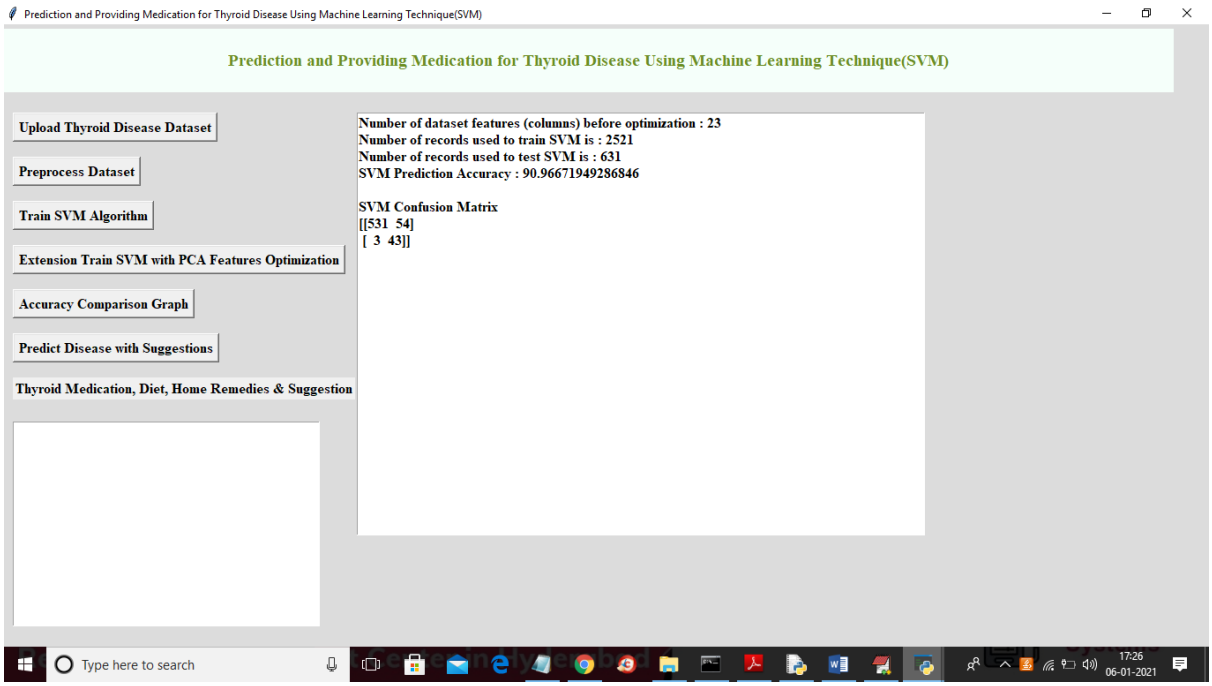



To upload a dataset related to thyroid disease, the aforementioned screen must be navigated to before the screen shown below can be accessed.

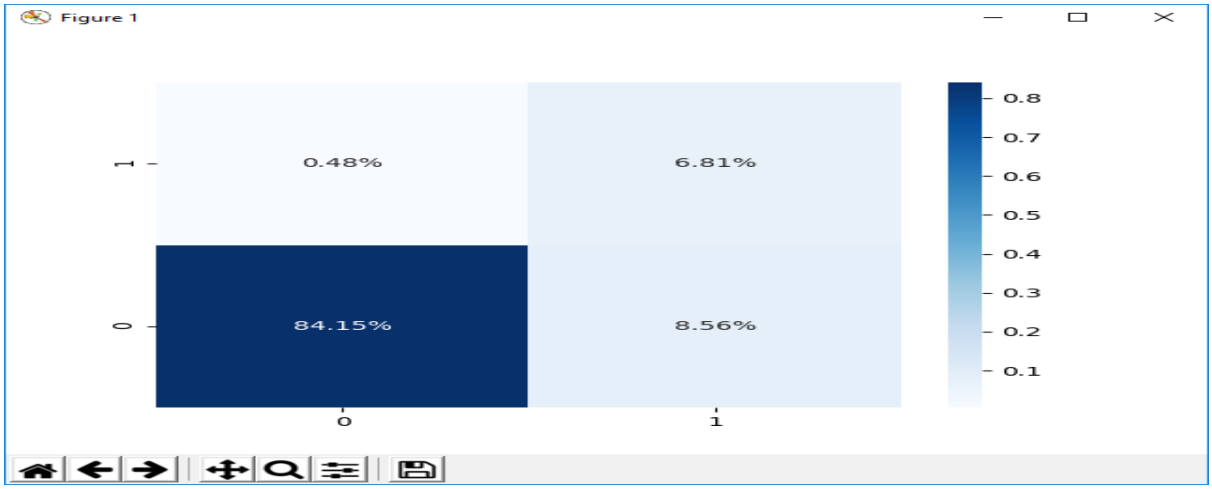



You can remove missing and NAN values from the dataset and split it into X and Y values, where X contains all dataset values and Y contains class label value, by

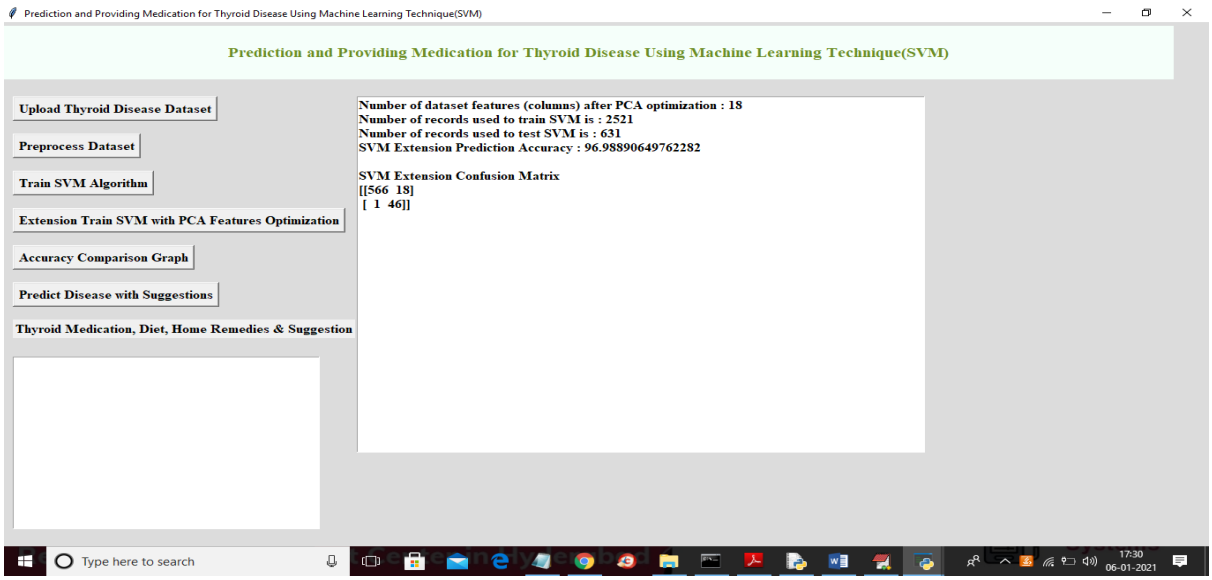
clicking the corresponding button in the screen shot above, after which the dataset will be loaded and a few records from it will be displayed..



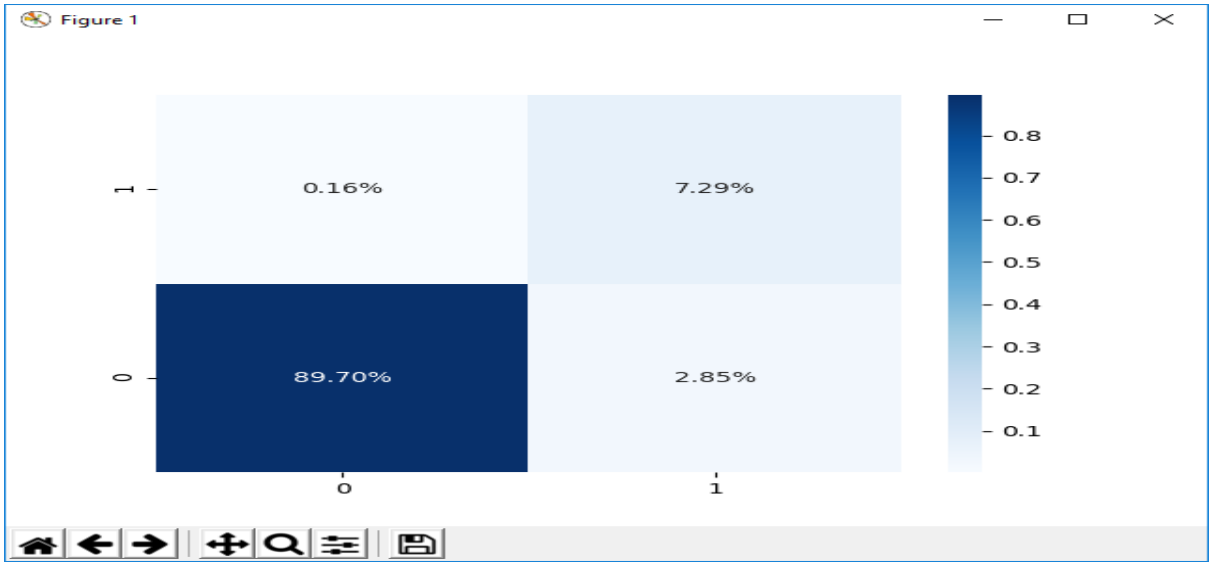

Above, we can see that the dataset has a total of 23 columns, that the SVM algorithm was trained on 2521 records using 631 test records, and that the prediction accuracy was 90.96% using a standard SVM. Furthermore, the application displays a confusion matrix of true and false prediction values, with 531 and 3 representing the correct prediction and 54 and 43 representing the incorrect one, respectively, and a graph representation of the matrix being provided below.



Click the "Extension Train SVM with PCA Features Optimization" button to train the SVM with PCA features optimization and obtain the prediction accuracy shown below the graph (84.15% and 6.81%).

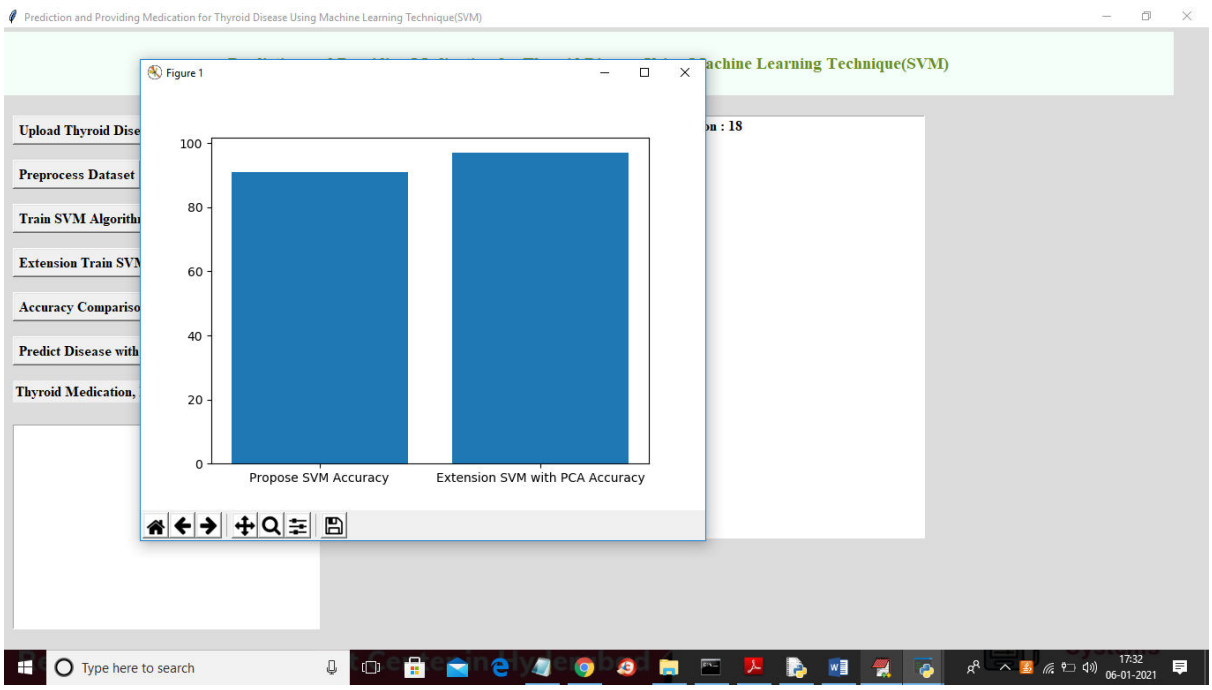



The PCA extension to the SVM shown above achieved a prediction accuracy of 96.08%, and the confusion matrix values it produced were significantly lower than those of standard SVM.

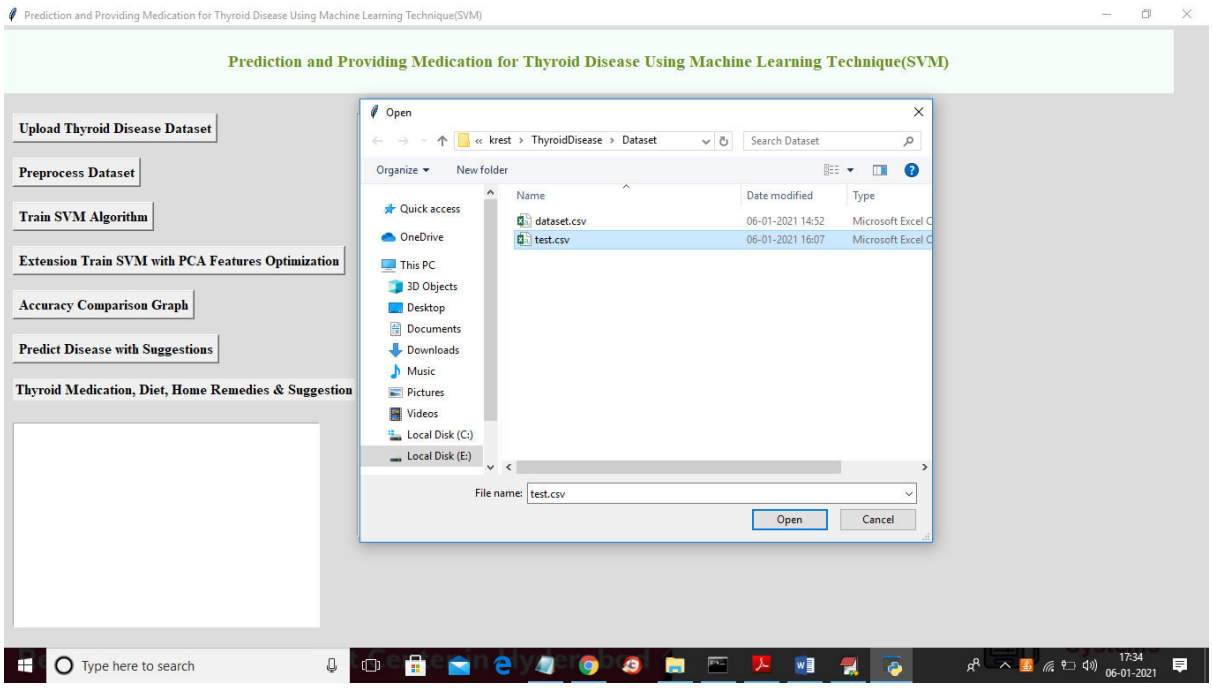



The correct prediction in the above graph is 89.70, and the other values are false predictions. To view the below accuracy comparison graph, click the 'Accuracy

Comparison Graph' button.

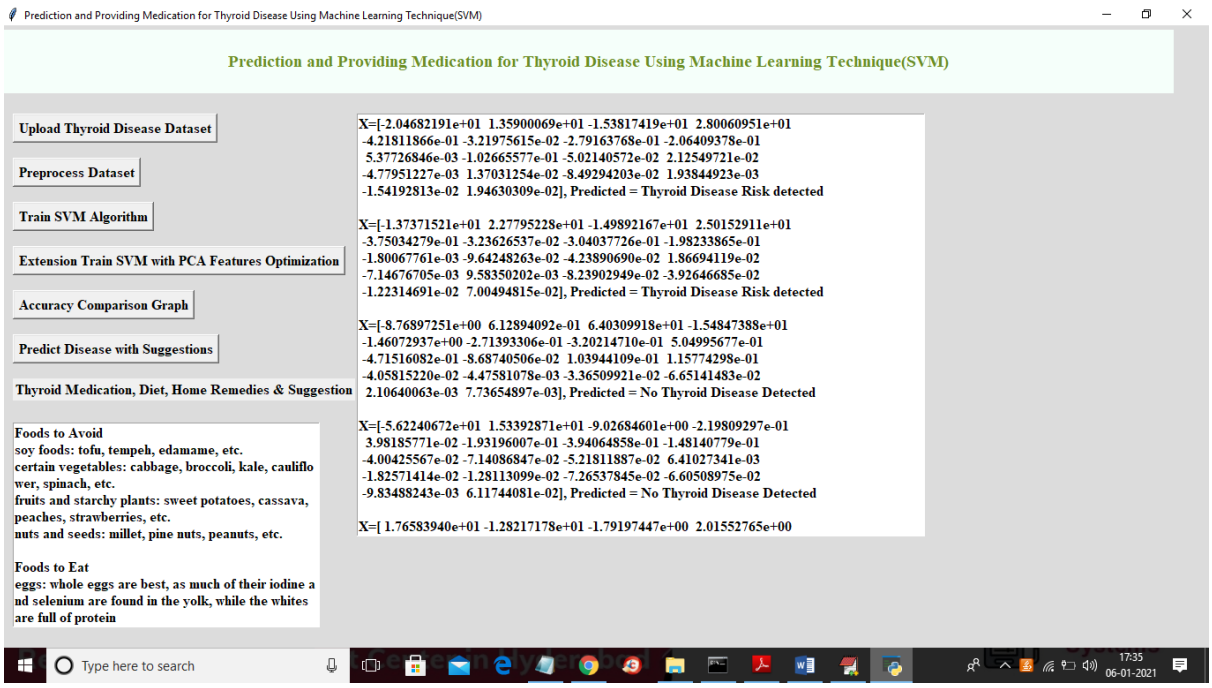



To upload new test data and predict if the data contains thyroid or not, click on the "Predict Disease with Suggestions" button; the x-axis of the above graph represents the name of the algorithm, and the y-axis represents the accuracy of the algorithm. Extension SVM with PCA is superior to normal SVM, as shown in the above graph.




To upload a test dataset and make a disease prediction, follow the steps outlined above, selecting and uploading the 'test.csv' file, and then clicking on the 'Open'

button.




Each record's test value is displayed in brackets above; following the brackets is information about whether or not a thyroid risk has been detected, and if one has been, a suggested diet and medication schedule appear in the left box.

## 6. CONCLUSION

In addition, this paper investigates the novel machine learning approaches that can be used to spot thyroid disorders. Many convenient analyses have been created and used in recent years to diagnose thyroid illness correctly and expertly. Based on the research conducted, it is clear that the two papers use different technologies with varying degrees of accuracy. The majority of research articles conclude that neural networks are superior than alternative methods. While there is no doubt that medical professionals everywhere have made enormous strides in their ability to identify thyroid issues, it is advised that people use a smaller subset of the available diagnostic criteria. Having more distinguishing features necessitates more comprehensive, time-consuming, and expensive health assessments.

When compared to other classifiers, we find that RFE provides the highest level of

accuracy when used as a feature selection strategy. Through the use of a real-time data set, we found that RFE considerably aids in the primary stage prediction of hypothyroidism. Since the outbreak, data collection has been extremely challenging. Thus, we have only 519 records because of this. As a result, we were unable to investigate on a more extensive data set due to the nature of the circumstance and the limitations imposed upon us. From what we were able to ascertain, prior research into the thyroid in Bangladesh is lacking. The information we have at our disposal is limited. We hope that more people in our country will get interested in working on this condition in the future so that we can use a more comprehensive dataset to improve our ability to anticipate the onset of disease in its basic stages. With any luck, it will assist the citizens of our country in preserving a thriving culture. To help patients save time and money, we need to create algorithms and predictive models of thyroid disease that reduce the number of criteria a doctor needs to meet to make a diagnosis.

## 7. FUTURE SCOPE

Further research can be conducted by applying image processing of ultrasonic scanning of thyroid pictures to anticipate thyroid nodules and cancer that are not detectable in blood test results. Thyroid disease prediction can encompass all thyroid-related disorders by integrating both data.

## REFERENCES

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

[2] YongFeng Wang,(2020). “Comparison Study of Radiomics and Deep-Learning Based Methods for Thyroid Nodules Classification using Ultrasound Images” published on IEEEAccess.

[3] Sunila Godara,(2018). “Prediction of Thyroid Disease Using Machine Learning Techniques” published on IJEE.

[4] Hitesh Garg,(2013). “Segmentation of Thyroid Gland

in Ultrasound image using Neural Network" published on IEEE.

[5] L. Ozyılmaz and T. Yıldırım,(2002). "Diagnosis of thyroid disease using artificial neural network methods," in: Proceedings of ICONIP'02 9th international conference on neural information processing (Singapore: Orchid Country Club, pp. 2033–2036).

[6] K. Polat, S. Sahan and S. Gunes,(2007) "A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted preprocessing for thyroid disease diagnosis," Expert Systems with Applications,(vol. 32, pp. 1141-1147).

[7] F. Saiti, A. A. Naini, M. A. Shoorehdeli, and M. Teshnehlab,(2009) "Thyroid Disease Diagnosis Based on Genetic Algorithms Using PNN and SVM," in 3rd International Conference on Bioinformatics and BiomedicalEngineering. ICBBE 2009.

[8] G. Zhang, L.V. Berardi,(2007) "An investigation of neural networks in thyroid function diagnosis," Health Care Management Science,1998, (pp. 29-37.)

[9] V. Vapnik,(2012).Estimation of Dependences Based on Empirical Data, Springer, New York.

[10] Obermeyer Z,(2016). Emanuel EJ. Predicting the future— big data, machine learning, and clinical medicine. N Engl ; (375:12161219).

[11] Breiman L.(2001) StatisticalModeling: the two cultures.Stat Sci. ;16:199-231..

[12] Ehrenstein V, Nielsen H, Pedersen AB, Johnsen SP, Pedersen L. (2017) Clinical epidemiology in the era of big data: new opportunities, familiar challenges. Clin Epidemiol. ; 9:245-250

[13] S. Godara and R. Singh,(2016) "Evaluation of Predictive Machine Learning Techniques as Expert Systems in Medical Diagnosis", Indian Journal of Science and Technology, (Vol. 910).

[14] Sunila, Rishipal Singh and Sanjeev Kumar.(2016) "A Novel Weighted Class based Clustering for Medical Diagnostic Interface." Indian Journal of Science and Technology (Vol 9).