Predicting Thyroid dysfunction with machine learning models

**Abstract.** The prevalence of thyroid dysfunction is now alarming worldwide, particularly in Africa. The treatment is effective only when detected and diagnosed accurately at early stages. The diagnosis of thyroid dysfunction requires good knowledge of various blood tests, which health practitioners usually interpret manually. A better option is in data mining which offers lot of classification techniques to automatically predict diseases based on patients’ data in hospitals and clinics. Machine learning, a branch of artificial intelligence employs different statistical and optimization operating rules that let the computer to “learn” from earlier cases and to detect hard-to-discern patterns from bulky, noisy or complex datasets. The drive for this research is to discern the most efficient model to predict thyroid dysfunction at early stages. In this research, the datasets used were obtained from the University of California, Irvine (UCI) data repository which contain 3774 patients records. These records include levels of free triiodothyronine (FT3), thyroid stimulating hormone (TSH), triiodothyronine (T3) and thyroxine (T4) amongst others. Support Vector Machine (SVM), decision tree and logistic regression machine learning models were implemented, their accuracy, precision, and recall were then compared. The best accuracy (0.981), precision (0.827) and recall (0.727), were obtained in the decision tree model.

**Keywords:** Thyroid dysfunction, machine learning, decision trees.

1. Introduction

In developing and densely populated countries, dietary iodine deficiency is the major determinant of thyroid pathology, resulting in a spectrum of iodine deficiency disorders, including goitres, hypothyroidism and mental retardation (Okosieme, 2006). In areas where the daily iodine intake is <50 μg, goitre is usually endemic, and when the daily intake falls <25 μg, congenital hypothyroidism is seen. The prevalence of goitre in areas of severe iodine deficiency can be as high as 80% (Vanderpump, 2011). Majority of these cases if undetected and diagnosed early can eventually lead to complications and death. In a world where automation and hospital management system are now being used in diagnosis of ailments and diseases, Thyroid dysfunction can also be easily detected from patients’ records via Machine Learning (ML).

Machine Learning algorithms have been used to achieve great advancements in disease prediction. This is due to its ability to predict the occurrence of diseases through classification techniques that classify the data (patients’ record) into predetermined categories. Rajam et.al. (2016) surveyed the various data mining techniques that can be used in thyroid dysfunction diagnosis. They suggested the usage of algorithms such as Naïve bayes, decision tree, back propagation and Support vector machine. Lui and Pappas (2015) used exploratory approach to compare different models in determining the presence of hyperthyroidism or hypothyroidism in previously undiagnosed patients who were presumed healthy. They concluded that thyroid dysfunction can be predicted to good accuracy using TSH, FTI, and TT4 in a simple decision tree model, but also recommended that (i) more recent datasets should be used, as the set used was from the mid-1980s; (ii) more complete data is needed, to see if other classification methods can perform better than decision trees; and (iii) a larger set of data from a larger variety of sources should be obtained.

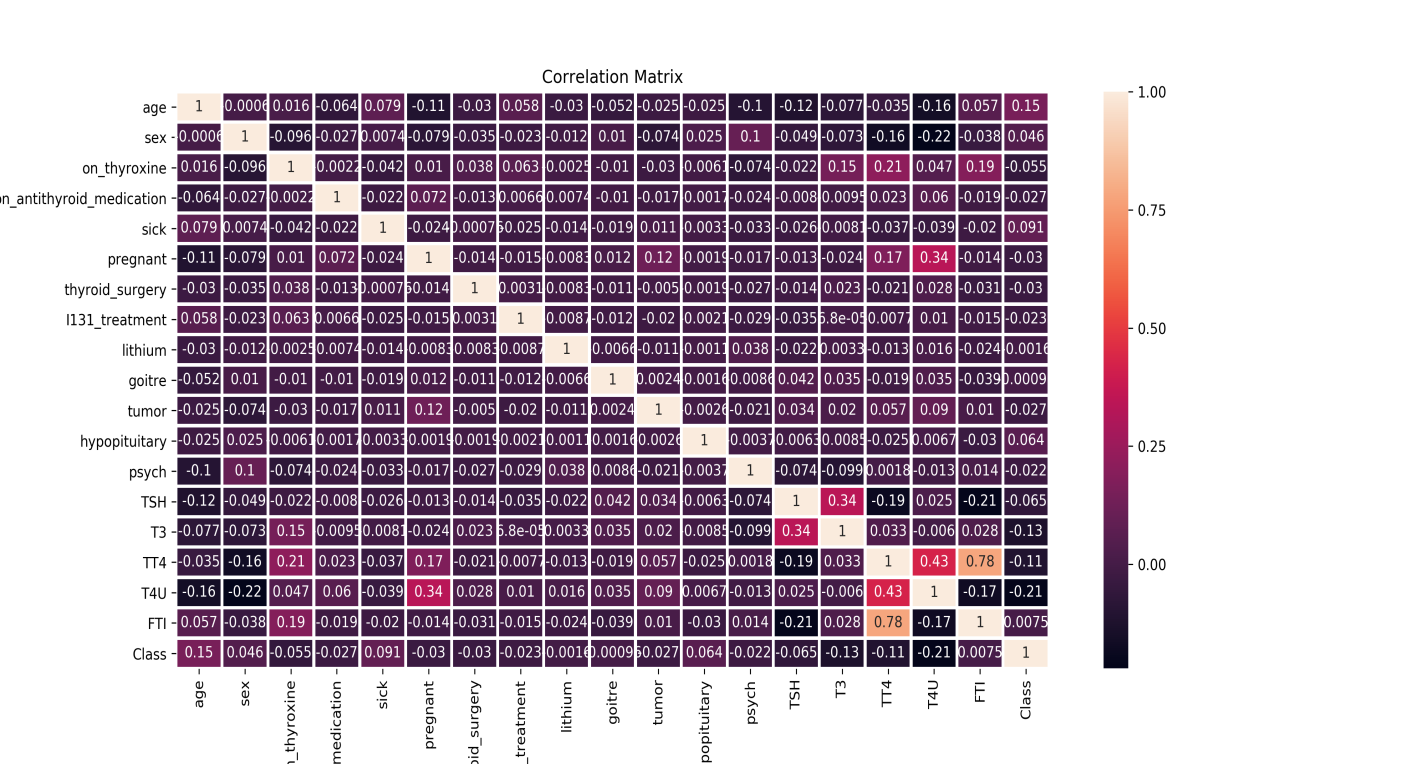
Based on the recommendations of Lui and Pappas (2015), this research employed a structured analytical method (Machine learning) to design predictive models. Modern datasets, whose features are more complete and large, were used in the prediction of thyroid dysfunction at its early stages. Three (3) models designed with different ML algorithms and these models were there compared.

1. Methodology

Dataset

In this study, the dataset used was obtained from University of California, Irvine (UCI) machine learning data repository (<http://archive.ics.uci.edu/ml/datasets/thyroid+disease>). The database contains 3774 patients’ data with features such as: FTI, TSH, TT4, T4U levels, age, health status at the time of test (sick or not sick), gender and pregnancy status for women etc. The dataset mostly contained binary annotations such as presence of pregnancy, goiter and others diseases that can affect thyroid stimulation in the body. There were many missing values in the dataset, thus, data cleaning was performed to either replace the missing values with median or drop those with missing values. Also from the correlation plot, it was realized that there were some features that had insignificant correlation to the target variable. An example is the TGB, referral source and Iodine-131treatment. These features were removed to enable an optimum performance of the model. 80% and 20% of the datasets were used for the training and test respectively.

The frequently used patient’s features for the diagnosis of thyroid diseases are given in table 1. Researchers have selected one or more of these features as inputs variables to the thyroid dysfunction prediction model. While some researchers used four (4) or five (5) features, this research employed the usage of twenty-one (21) features of the twenty-four (24) features in the dataset. Less discriminatory features were eliminated based on their correlation plot to thyroid dysfunction leaving a subset of the original features that still retain sufficient information needed to discriminate well among the classes. The correlation plot in figure 1 shows the relationship between thyroid dysfunction against each of the features used. The features used are outlined in Table 1 below.



**Fig 1:** Correlation Plot of Class against features

**Table 1:** Features used for the models

|  |  |
| --- | --- |
| **Attributes** | **Value type** |
| Age | In years |
| Sick | F, T |
| Sex | M, F |
| Thyroid surgery | F, T |
| Pregnancy | F, T |
| Iodine-131 treatment | F, T |
| query hypothyroid | F, T |
| query hyperthyroid | F, T |
| Lithium | F, T |
| Goitre | F, T |
| Hypopituitary | F, T |
| Psych | F, T |
| Tumor | F,T |
| T3 measured | Continuous |
| T3 | F, T |
| FTI measured | Continuous |
| TSH measured | Continuous |
| T4U | F, T |
| T4U measured | Continuous |
| TT4 | F, T |
| TT4 measured | Continuous |

Tools used

In developing this model, python modules (scikit-learn, matplotlib, numpy and pandas) were used. Matplotlib python library was used for visualization. It was used it to visualize the correlation between the features (such as age, sex, pregnancy status etc) and the target (thyroid dysfunction). Numpy was used for mathematical computation (the mean, median and standard deviations of the numerical features contained in the dataset). Scikit-learn was used in building the models (logistic regression, gradient boosting and decision trees), of all the types of learning algorithms that can be applied for this kind of predictive analysis, supervised learning algorithms was used.

**Model pipeline**

Figure 2 outlines the pipeline (the mode of operation) for building the model used. The original datasets obtained was first preprocessed by cleaning the data i.e. removing the empty features or replacing the empty numerical features with the median. The preprocessed data was then divided into the training dataset and the testing dataset. The training dataset was used to train the models, after which the model was tested with the testing data to obtain a predicted output.

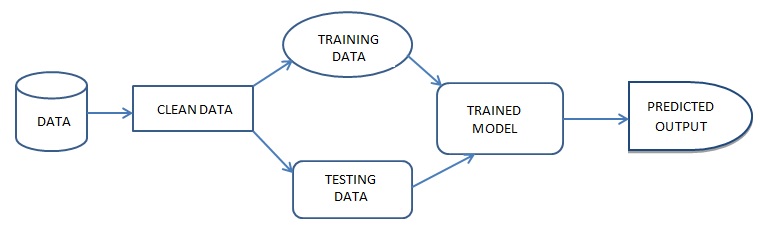


Fig 2: Model pipeline diagram

1. Results and Discussion

## Confusion Matrix Results of the models used for Thyroid Prediction

The total dataset in the database used is 3774. For training the models, 80% of the dataset was used (3019) while the remaining 20% (755) was employed in testing the models by supplying the test data to the classifier of decision trees, Support Vector Machines (SVM) and logistic regression algorithms. The prediction result obtained from each model is displayed in Table 2.

**Table 2:** Prediction results format

|  |  |  |
| --- | --- | --- |
| N = Test Data Size | Predicted NO | Predicted YES |
| Actual NO | TN | FP |
| Actual YES | FN | TP |

|  |  |  |  |
| --- | --- | --- | --- |
| DECISION TREE | N = 755 | Predicted NO | Predicted YES |
| Actual NO | 714 | 3 |
| Actual YES | 12 | 26 |

|  |  |  |  |
| --- | --- | --- | --- |
| LOGISTIC REGRESSION | N = 755 | Predicted NO | Predicted YES |
| Actual NO | 708 | 9 |
| Actual YES | 13 | 25 |

|  |  |  |  |
| --- | --- | --- | --- |
| SUPPORT VECTOR MACHINE | N = 755 | Predicted NO | Predicted YES |
| Actual NO | 714 | 3 |
| Actual YES | 33 | 5 |

Table 2: Confusion Matrix Result of each model

### **Logistic regression model result**

The logistic regression model had predicted 38 patients to have thyroid disease while 717 patients did not have thyroid disease with an accuracy score of 97.09%. The precision and recall scores are represented in the figure below:

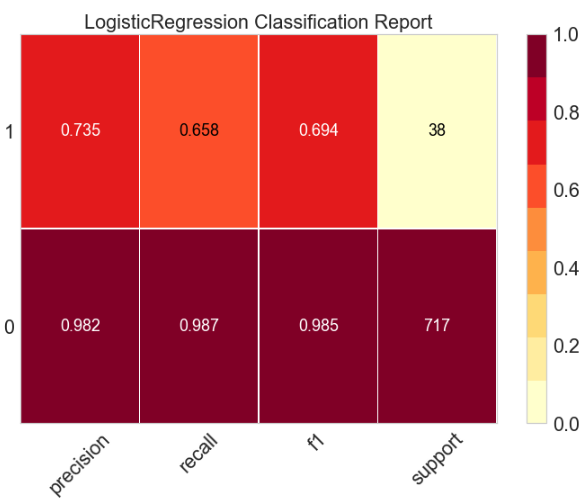


Fig 3: Classification report of Logistic Regression

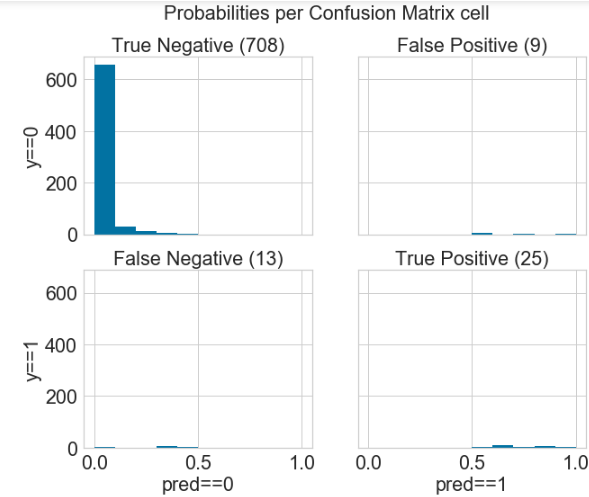
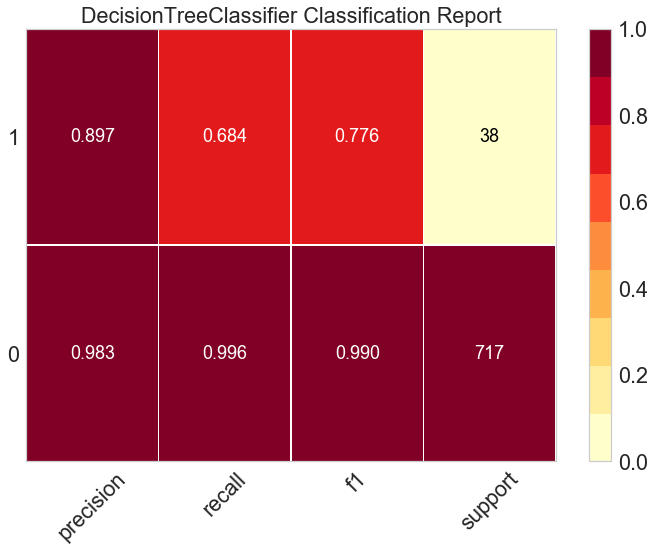


Fig 4: Confusion matrix of the logistic regression model

### **Decision tree model performance result**

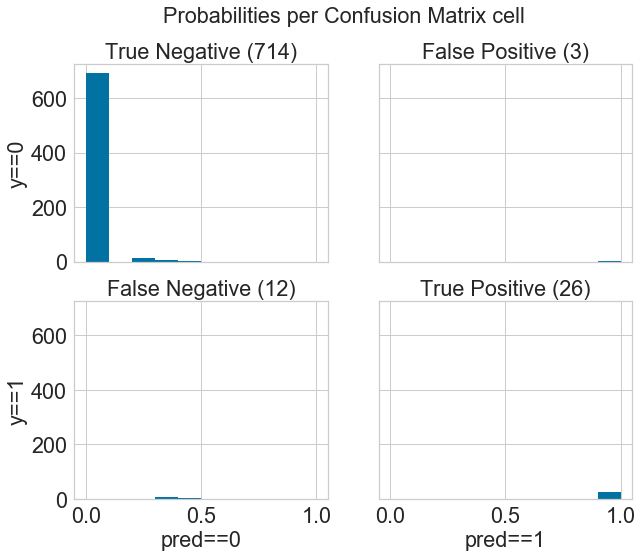
The decision tree model predicted that 38 people had thyroid disease while 717 people did not have thyroid disease with an accuracy of 98.01%. The precision and recall levels are shown in the figure below:



**Fig 5:** Classification report of Decision tree model

From the result shown in the figure above, the decision tree model predicts the positive label with a lower precision and recall values than it did predict the negative values with a higher precision. The logic behind is that the number of samples with positive labels are more than the number of samples with negative labels, so the model predicts the negative labels with a higher level of precision and accuracy than it does for the other levels.

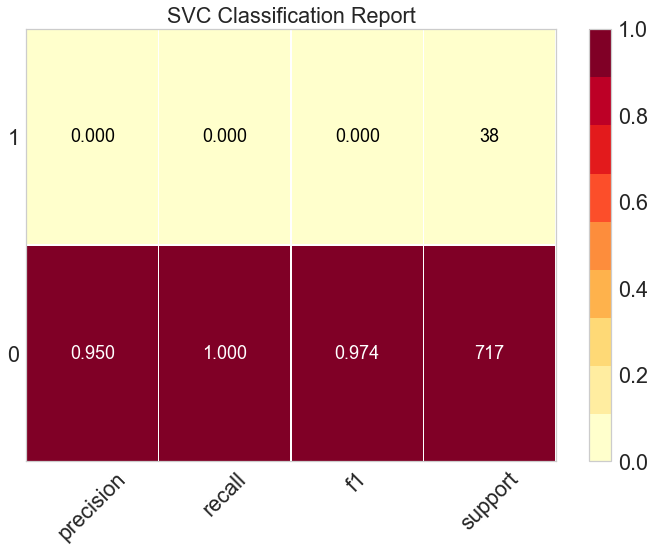
The probabilities per confusion matrix are shown in the figure below:



**Fig 6**: Confusion Matrix of decision tree model

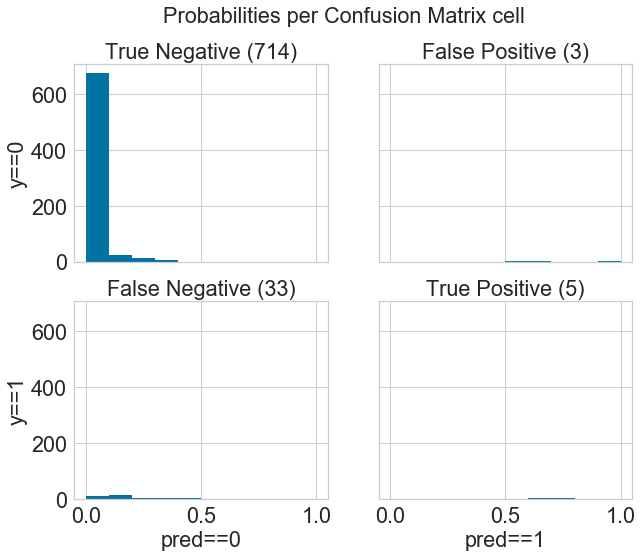
### **Support Vector Machine model result**

The support vector machine predicted that 38 people have thyroid disease with a precision of 0% and an accuracy of 94%, which is not too good for a model. The result is shown in the figure below:



**Fig 7:** Classification report of Support Vector Machine model

The confusion matrix of the support vector machine model is also shown below:



**Fig 8**: Confusion Matrix of Support Vector Machine Model

Of all the three models, the decision tree has the highest accuracy value of 98.01% which makes it the best model to use for the prediction of thyroid dysfunction from the data. The Support Vector Machine (SVM), has a lower accuracy score but a high recall and precision scores for predicting negative labels but a low precision and recall scores for predicting positive labels. This makes it not a very good model to use in the prediction of thyroid dysfunction. Logistic regression also performed fairly well with an accuracy of 97.09%.

Performance Evaluation and comparison

Based on the frequencies of TP, TN, FP and FN, the performance of each model was estimated in terms of Accuracy, Precision, Recall and False Prediction Rate. These performances are shown in Table 3.



where TP represents True Positive, TN means True Negative, FP False Positive and FN is False Negative

|  |  |  |  |
| --- | --- | --- | --- |
| **Models** | **Accuracy** | **Precision** | **Recall** |
| **Decision Trees** | 98.1 | 94.0 | 84.0 |
| **Logistic Regression** | 97.1 | 85.8 | 82.3 |
| **Support Vector Machine (SVM)** | 94.0 | 95.0 | 100 |

Table 3: Classification report of the three models used

**4. CONCLUSSION AND RECOMMENDATION**

There is no doubt that researchers worldwide have attained a lot of success to diagnose thyroid diseases but it is suggested to decrease the number of parameters used by the patients for diagnosis of thyroid diseases. More attributes means a patient has toundergo more number of clinical tests which is both cost effective as well time consuming. Thus there is need to develop such type of algorithms and thyroid disease predictive models which require minimum number of parameters of a person to diagnose thyroid disease and saves both money and time of the patient (Razia and Narasinga, 2016).

From the work carried out and the previous related work, we conclude that machine learning algorithms will perform better than human experts at predicting thyroid dysfunction will. Other research areas related to this work might include the use of more advanced techniques called deep learning to build predictive models. Deep learning technique even provides a higher accuracy than machine learning techniques and provides ability for the data to learn by itself from an unfamiliar and unknown data. This concept is otherwise known as Reinforcement Learning.

References

1. [Fontes R](https://www.ncbi.nlm.nih.gov/pubmed/?term=Fontes%20R%5BAuthor%5D&cauthor=true&cauthor_uid=24365659), [Coeli CR](https://www.ncbi.nlm.nih.gov/pubmed/?term=Coeli%20CR%5BAuthor%5D&cauthor=true&cauthor_uid=24365659), [Aguiar F](https://www.ncbi.nlm.nih.gov/pubmed/?term=Aguiar%20F%5BAuthor%5D&cauthor=true&cauthor_uid=24365659), [Vaisman M](https://www.ncbi.nlm.nih.gov/pubmed/?term=Vaisman%20M%5BAuthor%5D&cauthor=true&cauthor_uid=24365659). Reference interval of thyroid stimulating hormone and free thyroxine in a reference population over 60 years old and in very old subjects (over 80 years): comparison to young subjects. [Thyroid Res.](https://www.ncbi.nlm.nih.gov/pubmed/24365659) 2013 Dec 24; 6(1):13.
2. Gurmeet K, Er.Brahmaleen, Kaur S. Artificial Neural Networks for Diagnosis of Thyroid Disease. International Journal for Technological Research in Engineering.2014; 2 (1): 56-59.
3. Guttag JV. Introduction to Computation and Programming Using Python: With Application to Understanding Data. MIT Press. 2016. [*ISBN*](https://en.wikipedia.org/wiki/International_Standard_Book_Number) [*978-0-262-52962-4*](https://en.wikipedia.org/wiki/Special:BookSources/978-0-262-52962-4).
4. Jazzar MM. and Muhammad G. Feature selection based verification/identification system using fingerprints and palm print. Arabian J. Sci. Eng. 2013, 38 (4), 849–857.
5. Liu S, Liu S, Cai W, et al. Early diagnosis of Alzheimer’s disease with deep learning. In: International Symposium on BiomedicalImaging, Beijing, China. 2014:1015–18.
6. Lui AY, and Pappas AM. Thyroid Dysfunction: Prediction and Diagnostics. 2015. <https://docplayer.net/55037518-Thyroid-dysfunction-prediction-and-diagnostics.html>.
7. Mark P. J. Vanderpump; The epidemiology of thyroid disease, British Medical Bulletin, Volume 99, Issue 1, 1 September 2011, Pages 39–51, https://doi.org/10.1093/bmb/ldr030
8. Milmann KJ and Avaizis M. Scientiﬁc Python, volume 11 of Computing in Science & Engineering. IEEE/AIP, March 2011
9. Murphy, KP. Machine Learning - A Probabilistic Perspective. The MIT Press. 2012; pp. 245pp. [*ISBN*](https://en.wikipedia.org/wiki/International_Standard_Book_Number) [*978-0-262-01802-9*](https://en.wikipedia.org/wiki/Special:BookSources/978-0-262-01802-9).
10. Okosieme O.E. (2006), Impact of iodination on thyroid pathology in Africa. Journal of the Royal Society of Medicine, 99(8), 396-401.
11. Rajam K, Jemin R, Priyadarsini A. A Survey on Diagnosis of Thyroid Disease Using Data Mining Techniques. IJCSMC, Vol. 5, Issue. 5, May 2016, pg.354–358.
12. Rougier NP. Scientific visualization and matplotlib tutorial. Euroscipy 2012 & 2013. Available: [http://www.loria.fr/~rougier/teaching/matplotlib/matplotlib.html. Accessed 12 August 2014](http://www.loria.fr/~rougier/teaching/matplotlib/matplotlib.html.%20Accessed%2012%20August%202014).
13. Smith MR. and Martinez T. ["Improving Classification Accuracy by Identifying and Removing Instances that Should Be Misclassified"](https://dx.doi.org/10.1109/IJCNN.2011.6033571). Proceedings of International Joint Conference on Neural Networks (IJCNN 2011). pp. 2690–2697.
14. Srinivasan B and K. Pavya Diagnosis of Thyroid Disease Using Data Mining Techniques: A Study. International Research Journal of Engineering and Technology (IRJET), 2016, Volume: 03 Issue: 11; 1191-1194.