Thyroid Disease Detection Using Machine Learning and Pycaret

Dr. Tejashree T. Moharekar, Mr. Parashuram S. Vadar, Dr. Urmila R. Pol,
Dr. Pradip C. Bhaskar, Mr. Tejpal J. Moharekar

Assistant Professor, Department of Mathematics, Shivaji University, Kolhapur
tejashreemoharekar24@gmail.com
Research Student, Department of Computer Science, Shivaji University, Kolhapur
parashuramvadar@gmail.com
Assistant Professor, Department of Computer, Science, Shivaji University, Kolhapur
urmilec@gmail.com
Associate Professor, Department of Technology, Shivaji University, Kolhapur
pcb_tech@unishivaji.ac.in
Assistant Professor, Department of Commerce, Shri Shahaji Chhatrapati Mahavidyalaya, Kolhapur
tejpal1891@gmail.com

ABSTRACT

This work focuses on the analysis and classification models used in the prediction of thyroid disease, using data obtained from the UCI machine learning repository. Machine learning plays a crucial part in the process of disease prediction. In order to better predict the disease based on the parameters obtained from the dataset, this study applies multiple machine learning methods, including decision tree, random forest algorithm, KNN, and Naive Bayes, to the dataset for comparative comparison. The dataset was also modified to improve classification prediction accuracy. The proposed system uses Pycaret to apply various ML algorithms to the dataset, and then compares the results to improve the accuracy with which diseases can be predicted. The Naive Bayes classifier is the most accurate of these, at 95.91 percent.

Keywords: machine learning, thyroid, disease, prediction, hypothyroidism, classification, pycaret

I. INTRODUCTION

Diseases of the thyroid are extremely widespread across the globe. Thyroid diseases place a tremendous strain not just in the United States but also in India. It has been predicted that approximately 42 million individuals in India suffer from thyroid illnesses, which is a figure derived from a number of studies that have been conducted on thyroid disease. In recent years, there has been a proliferation in the accessibility of thyroid function testing, which has resulted in an increase in the number of people exhibiting symptoms that could be related to either hypothyroidism or hyperthyroidism being tested for the condition. Congenital hypothyroidism is likely the most significant of the several types of hypothyroidism because it necessitates an early diagnosis, which is then typically followed by adequate treatment that can delay or prevent the development of brain impairment. The field of medicine has benefited from the development of computational biology in recent years. It made it possible to collect the patient records that had been preserved for the purpose of medical disease prediction. In order to make an accurate diagnosis of the disease in its early stages, various sophisticated prediction algorithms are currently on the market.

Although the medical information system contains a large number of data sets, it does not have any intelligent systems that are able to quickly analyse diseases. Over the course of time, machine learning algorithms play an important part in the process of establishing a prediction model by helping to solve the complex and nonlinear challenges that arise during this process. Any disease prediction models need to give vital importance to the features that may be selected from the various datasets, and those features need to be able to be easily employed as a classification in healthy patients. If this is not avoided, a healthy patient may be subjected to unneeded treatment because they were incorrectly classified. Because of this, the factuality of predicting any disease in connection with thyroid disease carries the highest cardinality.

II. LITERATURE REVIEW

The authors suggest a method for predicting Hashimoto's thyroiditis (primary hypothyroid), binding protein (increased binding protein), autoimmune thyroiditis (compensated hypothyroid), and non-thyroidal syndrome (NTIS) (concurrent non-thyroidal illness). Extensive trials demonstrate that the extra tree classifier-based selected feature produces the best results with 0.99 accuracy and an F1 score when utilised in conjunction with the random forest classifier. Results indicate that machine learning models are a superior option for detecting thyroid illness in terms of accuracy and computational complexity (Rajasekhar Chaganti & Ashraf, 2022).

The study illustrated how to employ logistic regression, decision trees, and kNN as a classification tool, as well as provided insight into how to anticipate thyroid disease. UC Irvin's knowledge discovery in databases archive has been used by the machine learning repository's thyroid data set (Gyanendra Chaubey, 2021).

Authors proposed a thyroid disease prediction model using K-Nearest Neighbor (KNN), Naive Bayes, and Decision Trees machine learning classification methods. Experiments are conducted using thyroid data from the machine learning repository at UCI. The dataset consists of normal, hypothyroid, and hyperthyroid classes. Several criteria, including Accuracy, Precision, F-Measure, and Recall, are used to evaluate the performance of the three methods via 10-fold cross-validation. In a classification task involving three classes of thyroid illnesses, the decision tree was 99.7% more accurate than both Nave Bayes and KNN (Peya, Chumki, & Zaman, 2021).

Eleven different machine learning algorithms were evaluated by the authors of the study to discover which one is superior in terms of its ability to reliably predict thyroid risk. This study makes use of the dataset known as Sick-euthyroid, which was obtained from the machine learning repository at the University of California, Irvine.

The accuracy score does not provide a reliable indication of the results of the prediction in this dataset because the majority of the target variable classes are one. As a result, the metrics for evaluation include ratings for both accuracy and recall. In addition, the F1-score generates a single number that strikes a healthy balance between precision and recall if an uneven distribution class is present. Because it is one of the most effective output measurements for unbalanced classification issues, the F1-score is finally used to evaluate the effectiveness of the machine learning algorithms that have been deployed. In terms of accuracy, the experiment demonstrates that the ANN Classifier, which has a score of 0.957 on the F1 scale, surpasses the other nine algorithms (Islam SS, 2022).

The purpose of this study is to make a prognosis on the trend of LT4 treatment for patients who suffer from hypothyroidism. In order to achieve this goal, a specific dataset that contains medical information relating to patients currently being treated in the hospital known as "AOU Federico II" in Naples was developed. Because the clinical history of each patient is available over time, it was possible to predict the course of treatment for each patient based on the trend of the hormonal parameters and the other attributes that were taken into consideration. This enabled the researchers to determine whether the dosage of the patient's medication should be increased or decreased. The authors made use of a variety of machine learning methods when carrying out the investigation. In particular, they examined the differences and similarities between the outcomes of ten distinct classifiers. The performance of the various algorithms demonstrates positive results, particularly in the instance of the Extra-Tree Classifier, where the accuracy approaches 84% (Aversano, et al., 2021).

The authors used data from Iraqi persons, some of whom have an overactive thyroid gland and others who have hypothyroidism, to classify cases of thyroid disease into the three categories of hyperthyroidism, hypothyroidism, and normal. Linear discriminant analysis, k-nearest neighbors, multilayer perceptron (MLP), naive bayes, support vector machines, random forests, decision trees, and naive bayes (salman & Sonuç, 2021).

The thyroid gland can also be the location of different kinds of tumors and can be a dangerous place where endogenous antibodies wreak havoc (autoantibodies) (a. c.c.Heuck, 2000).

Doctors agree that early identification, diagnosis, and treatment are crucial in halting the spread of disease and saving lives. Early detection and differential diagnosis improve treatment outcomes for several types of abnormalities. Clinical diagnosis is often seen as challenging, despite several efforts to improve the process (Kouroua, 2015).

III. RESEARCH METHODOLOGY

Data Collection Machine learning algorithms are utilized in the speedy and early identification of thyroid illnesses and other diseases, since they have gained prominence in the medical profession and aid in disease diagnosis and classification. The dataset utilised in the study was obtained from the UCI machine learning repository.

As the data obtained consist of 29 variables or attributes where all the attributes were taken in our study which consist of 'age', 'sex', 'on thyroxine', 'query on thyroxine', 'on antithyroid medication', 'sick', 'pregnant', 'thyroid surgery', 'I131 treatment', 'query hypothyroid', 'query hyperthyroid', 'lithium', 'goitre', 'tumor', 'hypopituitary', 'psych', 'TSH measured', 'TSH', 'T3 measured', 'T3', 'TT4 measured', 'TT4', 'T4U measured', 'T4U', 'FTI measured', 'FTI', 'TBG measured', 'referral source', 'binaryClass' as target feature.

Data Preprocessing Pre-processing the data is a crucial stage in machine learning because it helps expose hidden information in the data through careful analysis and the identification of previously unknown patterns. Pre-processing entails a variety of steps, such as data cleaning, data preparation, etc. After identifying a set of missing data in this data, where the missing features were identified and worked to fill in, researchers were able to obtain the data in a good and better way, free from lost data, as the data became organized and good and free of any defect or problem so that one could work on it smoothly and well.

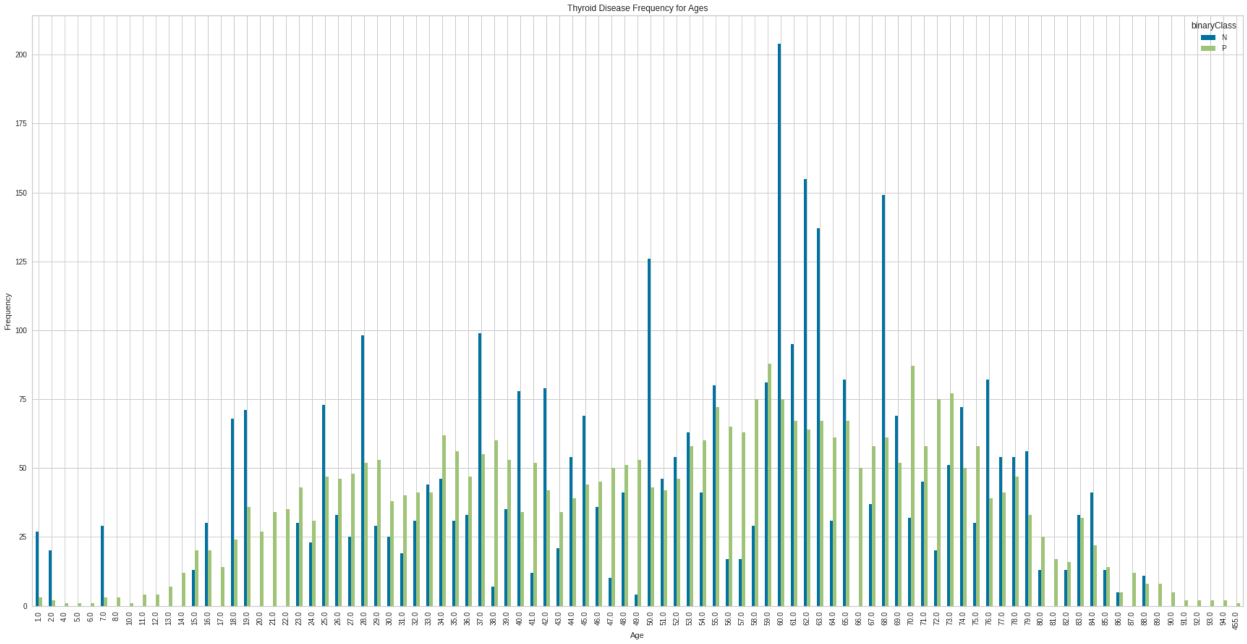
Data Machine Learning Techniques The major purpose of applying machine learning algorithms is to differentiate between thyroid illness. To get results from various machine learning methods applied to the dataset, Pycaret is employed. The classification feature in PyCaret is a supervised machine learning module used to categorise elements into a binary class using a variety of methods and algorithms.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

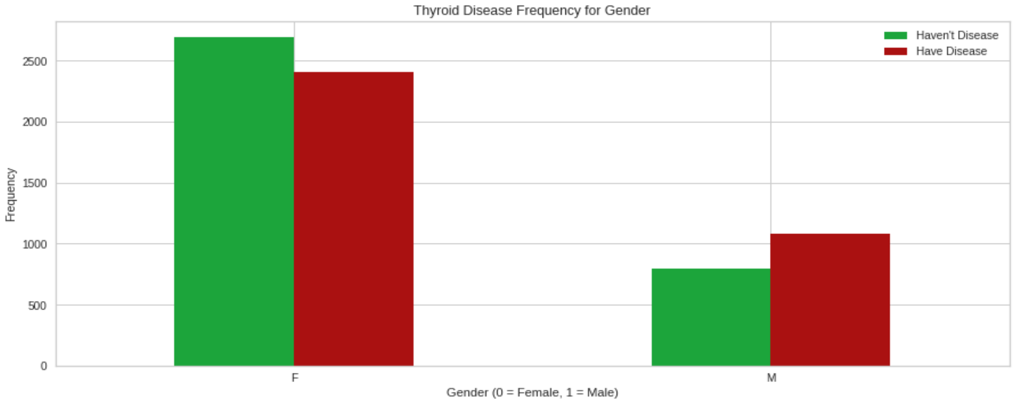
The correlation between thyroid health and lipid profiles varies with age and gender. Thyroid hormone secretion, metabolism, and action all undergo a number of shifts as we become older. Serum levels of thyroid stimulating hormone and T3 often drop with age, whereas serum levels of free T4 tend to remain stable.

A bigger proportion of the older population suffers from thyroid dysfunction than the younger population.

The graph below shows us the frequency of thyroid disease with ages.

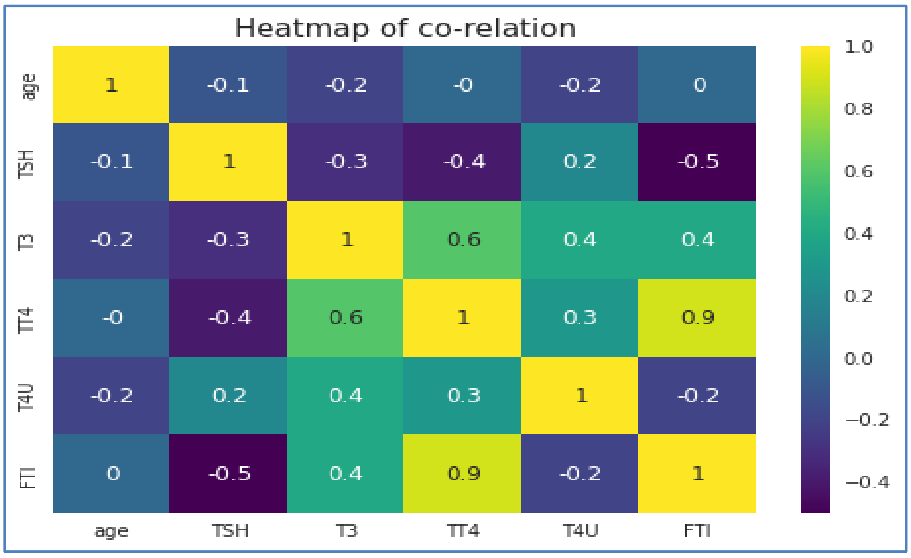


It was discovered that the prevalence of thyroid disease is significantly higher in females than in males. The incidence of thyroid dysfunction was found to be substantially higher in females than in males. The graph presents a comparison of males and females, those who have thyroid problems and those who do not.



The matrix illustrates the correlation between all of the many possible pair-wise combinations of values in the table. It is an effective tool for summing up enormous datasets, finding patterns in the data that has been provided, and visualizing those patterns. Correlation is helpful in getting basic information around which variables may be more important than the others for including variables that appear to have a reliable relationship with the output variable that researchers are looking to predict. This is true whether the task at hand is regression or classification.

According to the findings of the correlation heat map, the factors from the dataset, namely age, TSH, T3, TT4, T4U, and FTI, are more reliable to each other when attempting to forecast whether or not a person will have thyroid problems or not. It has been discovered that T3 and TT4 have a significant link with one another, while TT4 and FTI also have a significant correlation. Additionally, it was found that FTI and T3 had a relation to each other accordingly. The most common method for diagnosing hyperthyroidism, which is a condition in which the body produces an excessive amount of thyroid hormone, is by the use of a T3 test.

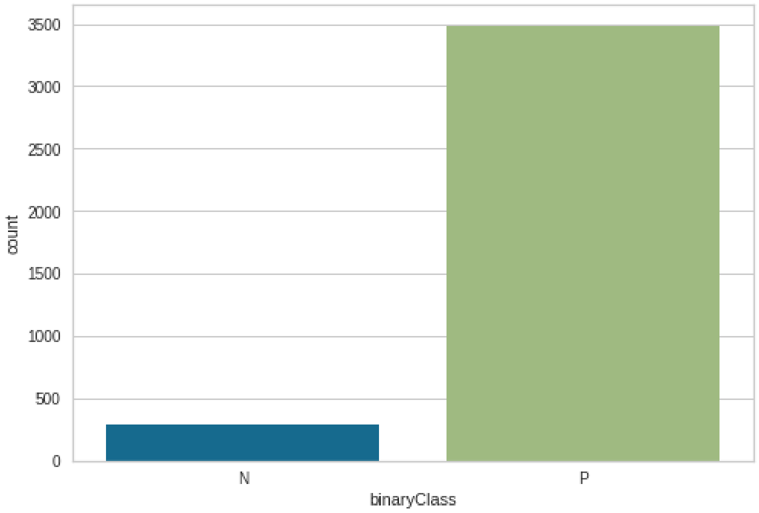


DESCRIPTIVE STATISTICS

The main purpose of descriptive statistics is to provide a brief summary of the samples and the measures done on the features which are numeric in nature. The following table shows the descriptive statistics of the features in the dataset.

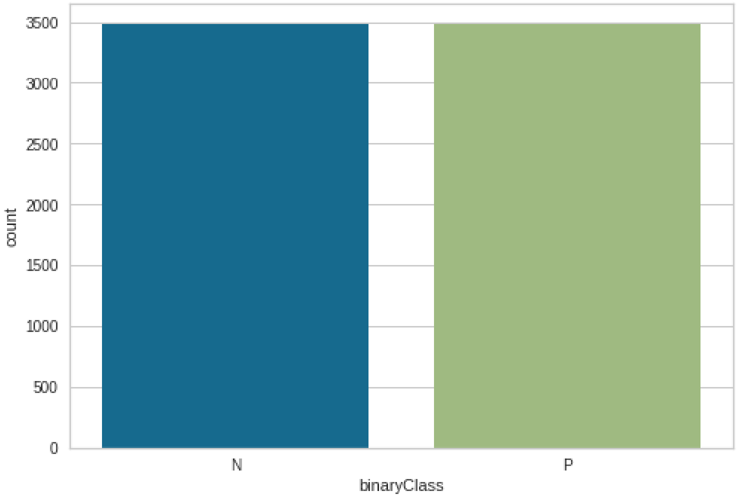
	age	TSH	T3	TT4	T4U	FTI
count	3772.000000	3772.000000	3772.000000	3772.000000	3772.000000	3772.000000
mean	51.737805	4.608713	2.010748	107.871103	0.994487	109.401034
std	20.082643	23.336066	0.738282	34.541135	0.185163	31.514921
min	1.000000	0.005000	0.050000	2.000000	0.250000	2.000000
25%	36.000000	0.200000	1.700000	89.000000	0.890000	94.000000
50%	54.000000	1.200000	2.000000	102.000000	0.990000	104.000000
75%	67.000000	2.425000	2.200000	123.000000	1.070000	121.250000
max	455.000000	530.000000	10.600000	430.000000	2.320000	395.000000

Imbalanced Data: Imbalanced Data refers to the types of datasets that have an unequal distribution of observations throughout the target class. This means that one class label will have a very high number of observations, while the other class label will have a very low number of observations. This is seen in the dataset that is being studied at the moment.



Imbalanced Data for Output Class

The Resampling (Oversampling) The method of resampling, also known as oversampling, is utilised in order to increase the representation of the minority group. When the dataset is unbalanced, we can use replacement to oversample the minority class in order to correct the imbalance.



Balanced Data for Output Class

The performance of our model may be more accurately evaluated with the use of a method called k-fold cross-validation. The model is evaluated with the help of various subsets of the data set that are used as the validation set. Our data set has been folded into K different groups. K is the number of folds that we wish to break your data into, and it reflects the number of folds. In this study, we employ 10-folds, which means that the data set is divided up into ten different portions.

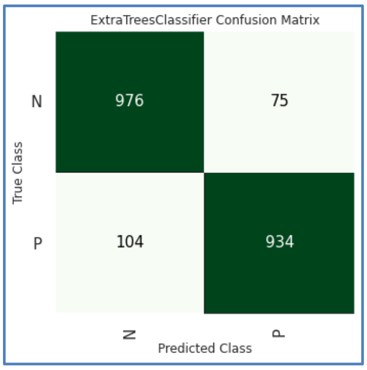
The results of various machine learning models that used the feature set are presented in the table below. Because of the selected feature, the data become more linearly separable. This makes it easier for the SVM to build a hyperplane with a good margin so that it can classify the data. As a result, the SVM's performance increases to 0.98. Due to the fact that a large feature set is necessary for a decent fit, the accuracy of the tree-based model ADA is reduced to 0.87, and the accuracy of the Gradient Boosting Classifier is also reduced to 0.87.

PERFORMANCE OF MACHINE LEARNING MODELS

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
Extra Trees Classifier	0.9922	0.9998	0.9865	0.9979	0.9922	0.9844	0.9845	2.831
SVM - Linear Kernel	0.9877	0.0000	0.9755	1.0000	0.9875	0.9754	0.9758	0.167

Random Forest Classifier	0.9871	0.9998	0.9763	0.9979	0.9869	0.9741	0.9744	2.467
Ridge Classifier	0.9867	0.0000	0.9746	0.9987	0.9865	0.9733	0.9736	0.158
Linear Discriminant Analysis	0.9863	0.9956	0.9726	1.0000	0.9861	0.9725	0.9729	2.064
Naive Bayes	0.9846	0.9847	0.9693	1.0000	0.9844	0.9692	0.9698	0.092
Logistic Regression	0.9842	0.9988	0.9697	0.9988	0.9840	0.9684	0.9689	1.533
Decision Tree Classifier	0.9750	0.9750	0.9501	1.0000	0.9743	0.9499	0.9512	0.867
Light Gradient Boosting Machine	0.9721	0.9972	0.9501	0.9941	0.9715	0.9442	0.9453	0.362
Gradient Boosting Classifier	0.8785	0.9579	0.7851	0.9663	0.8662	0.7571	0.7710	5.009
Ada Boost Classifier	0.8715	0.9268	0.7855	0.9495	0.8597	0.7432	0.7547	1.248
K Neighbors Classifier	0.8514	0.9569	0.7073	0.9949	0.8266	0.7031	0.7347	6.046
Quadratic Discriminant Analysis	0.8342	0.8337	1.0000	0.7518	0.8582	0.6681	0.7083	1.969
Dummy Classifier	0.5013	0.5000	1.0000	0.5013	0.6679	0.0000	0.0000	0.034

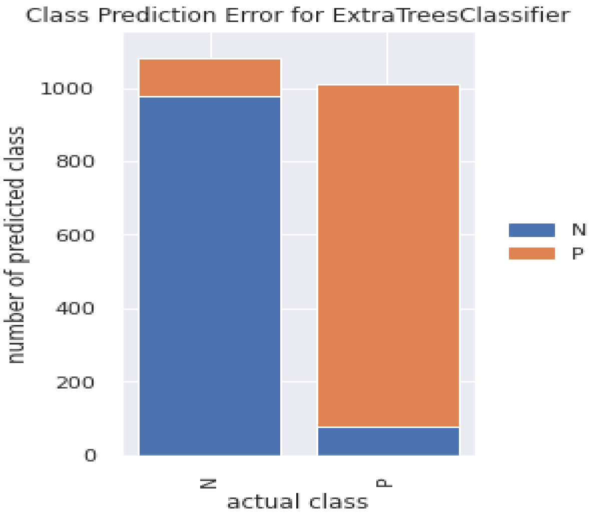
Confusion Matrix for Best Model: Extra Trees Classifier



Values predicted by the best model Extra Trees Classifier for the last five records from the dataset with the accuracy score is given below:

binaryClass	Label	Score
P	P	0.87
P	P	0.99
P	N	0.54
P	P	0.86
P	P	0.96

Error plot gives the errors and residual values from the best model.



V. CONCLUSION

Thyroid disease identification has arisen as an essential medical concern, and in order to solve it, efficient automatic prediction models are required. This problem has been alarmingly increasing over the past few years. Existing research concentrate their attention almost exclusively on model optimization and feature engineering, while feature selection receives far less attention. In addition, the dataset that was utilized for the evaluation of the model was obtained from the machine learning repository at UCI. Based on the findings, it appears that the additional tree classifier has the potential to produce an accuracy of 0.99. In a similar vein, the findings of a 10-fold cross-validation test support these conclusions. Out of the total 14 machine learning models top three models are Extra Trees Classifier with asn accuracy of 99.22% followed with SVM - Linear Kernel with an accuracy of 98.77% and Random Forest Classifier with an accuracy of 98.71% respectively. Cross validation allows for performance comparisons between several models that have been examined using Pycaret.

VI. REFERENCES

1. c.c.Heuck. (2000). Retrieved from World Health Organization: https://www.who.int/.

2. Aversano, L., Bernardia, M. L., Cimitileb, M., Iammarinoa, M., Macchiac, P. E., Nettorec, I. C., et al. (2021). Thyroid Disease Treatment prediction with machine learning approaches. Procedia Computer Science Elesvier , 1031-1040.

3. Gyanendra Chaubey, D. B. (2021). Thyroid Disease Prediction Using Machine Learning Approaches. Springer Nature , 233–238.

4. Islam SS, H. M. (2022). Application of machine learning algorithms to predict the thyroid disease risk: an experimental comparative study. PeerJ Comput Sci .

5. Kouroua, K. E. (2015). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal , 8–17. .

6. Peya, Z. J., Chumki, M. K., & Zaman, K. M. (2021). Predictive Analysis for Thyroid Diseases Diagnosis Using Machine Learning. International Conference on Science & Contemporary Technologies (ICSCT) (pp. 1-6). Dhaka, Bangladesh: IEEE.

7. Rajasekhar Chaganti, F. R., & Ashraf, I. (2022). Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques. Cancers MDPI , 1-23.

8. salman, K., & Sonuç, E. (2021). Thyroid Disease Classification Using Machine Learning. 2nd International Conference on Physics and Applied Sciences (ICPAS 2021) (pp. 1-12). Journal of Physics: Conference Series.