MACHINE LEARNING BASED CLASSIFICATION OF THYROID CONDITIONS

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology

In

Computer Science and Engineering

School of Engineering and Sciences

Submitted by

Sai Teja Kambhampati | AP20110010745

Hariprasad Chintham | AP20110010740

Naveen Krishna Bhogineni | AP20110010755

Under the Guidance of

Anuj Pradeep Deshpande

SRM University–AP

Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240

[DEC, 2022]

Certificate

Date: 13-DEC-22

This is to certify that the work present in this Project entitled "**MACHINE LEARNING BASED CLASSIFICATION OF THYROID CONDITIONS**" has been carried out by **K.Sai Teja, Ch.HariPrasad, B.Naveen Krishna** under my/our supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology in **School of Engineering and Sciences**.

Supervisor

(Signature)

Anuj Pradeep Deshpande

Assistant Professor ,ECE.

Acknowledgements

I would like to thank Anuj Pradeep Deshpande sir for providing us with the opportunity to work on this group project. We are grateful for their guidance and support throughout the project.

And also we would like to thank SRM University for providing us this project based learning.

Table of Contents

Abstract

The purpose of this project is to create a machine learning system that can accurately identify thyroid diseases based on patient medical data, such as their medical history, physical examination results, and lab tests. The system will be able to classify each patient's thyroid condition and provide accurate diagnosis. The system will also be capable of making decisions on whether to proceed with medical treatment or not. The proposed system will be based on supervised learning algorithms. The results of this project will be used to improve the accuracy and efficiency of diagnosis of thyroid diseases.

This project will benefit the medical field by streamlining the diagnosis process and reducing the time and cost associated with it. The system will also provide a more accurate diagnosis and reduce the rate of misdiagnosis. Additionally, this system can be used to reduce the number of unnecessary medical tests and treatments and improve the quality of care for patients.

Statement of Contributions

Naveen Krishna

I am responsible for the contribution of dataset, data preprocessing and model evaluation. And helped in writing the project report.

Sai Teja

I am responsible for feature selection, balancing the dataset and model implementation of this project. I collaborate with other team members to build a better classification system.

Hari Prasad

I assisted the team in the project by reading research papers, providing feedback and suggestions, and offering valuable insights. I also aided the group to resolve issues that arose during development and suggested solutions for them. Lastly, I wrote the project report with the help of my team.

Abbreviations

TSH - Thyroid stimulating hormone

FT4 - Free thyroxine

Tg - Thyroglobulin

TPO - Thyroid peroxidase

T3 - triiodothyronine

T4 - thyroxine

1. Introduction

The thyroid, a gland residing in the neck, is responsible for producing hormones that regulate metabolism, growth, and development. If it is not functioning properly, it can lead to a range of symptoms and issues, including fatigue, depression, and weight gain. Therefore, early diagnosis and treatment of thyroid disorders is vital in order to manage them effectively.

Hyperthyroidism is a medical condition in which the thyroid gland produces an excessive amount of thyroid hormones, leading to an imbalance in the body. [3].

In this report, we will discuss how machine learning can be used to detect thyroid disorders. We will explore how different machine learning algorithms, can be used to classify patients into two categories: those with thyroid disorders and those without. We will explore how data pre-processing and feature engineering techniques can be used to boost the accuracy of the model.

Hypothyroidism is a condition in which the production of thyroid hormones is reduced. Common signs and symptoms of hypothyroidism include obesity, low heart rate, increased temperature sensitivity, neck swelling, dry skin, hand numbness, hair issues, heavy menstrual cycles, and intestinal problems. If left untreated, these symptoms can worsen over time. [10]

Hypothyroidism is a medical condition in which the body produces a lower than normal amount of thyroid hormones. If not treated, the symptoms of hypothyroidism can worsen over time and include obesity, low heart rate, increased sensitivity to cold, neck swelling, digestive issues, numbness in the hands, hair loss, dry skin, and heavier menstrual cycles.

Overall, machine learning can be used to develop personalized treatment plans for patients with thyroid disorders. By analyzing patient data, machine learning algorithms can learn to identify patterns in the data that can be used to create individualized treatments that are tailored to each patient's needs. This can lead to improved outcomes, better patient compliance, and greater cost savings.

2. Methodology

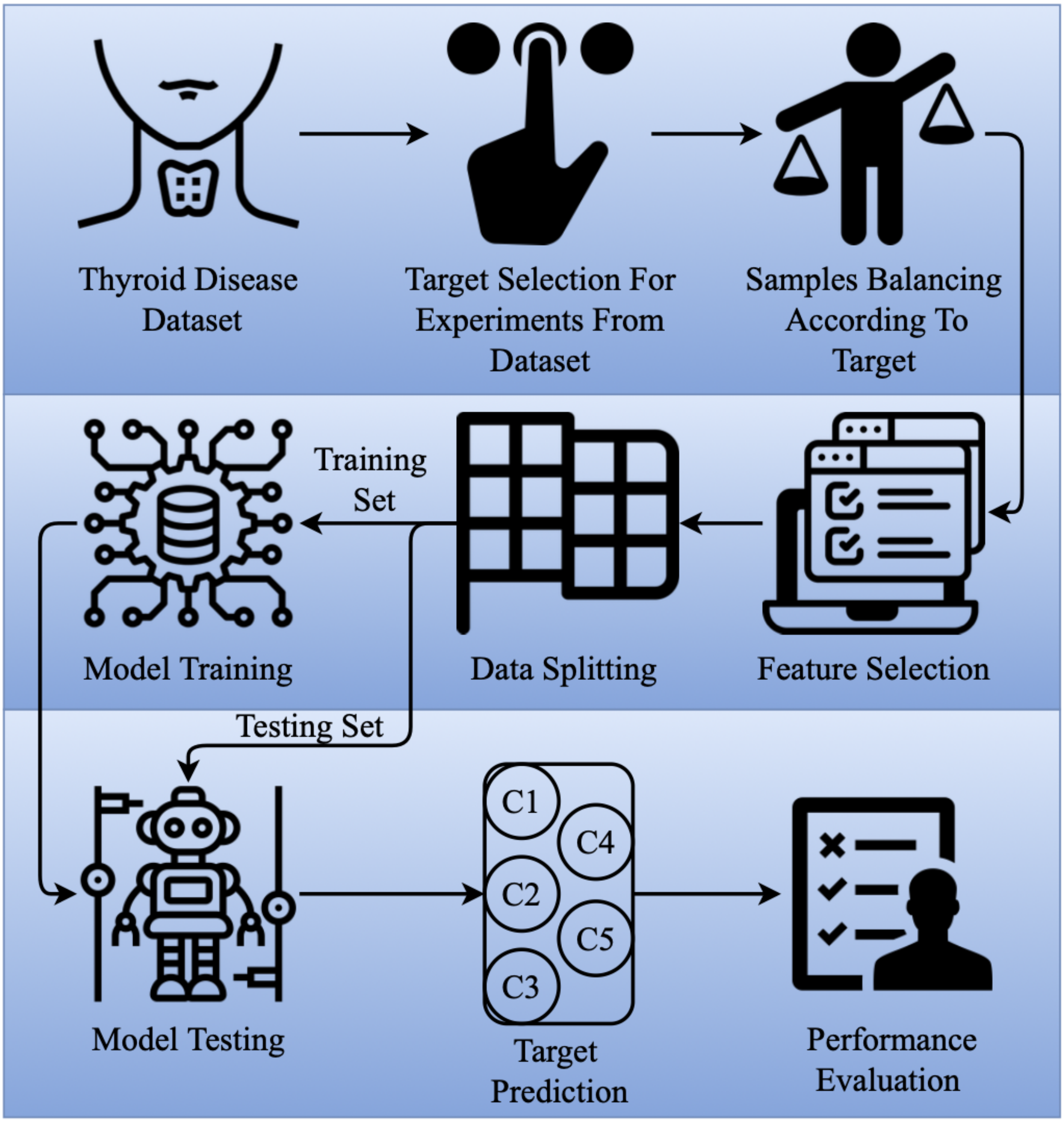


FIG: Flow chart for methodology

2.1. Data Preprocessing :

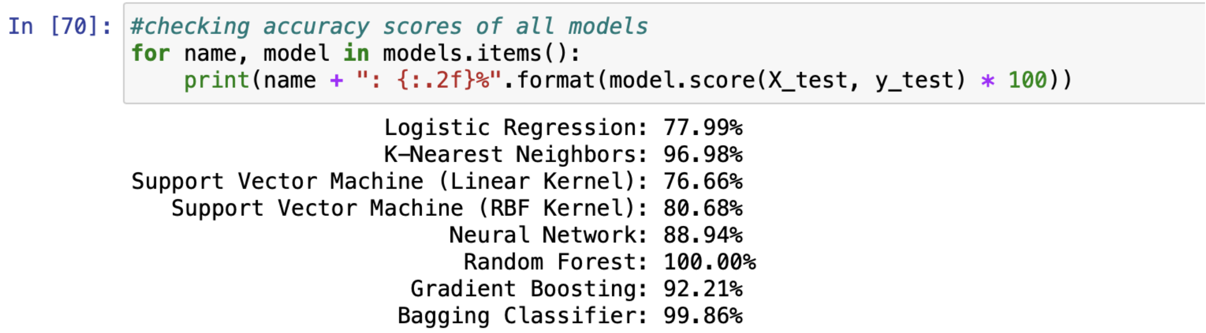
The first step in this methodology is to preprocess the dataset and check for any missing values or any invalid values. In this project, missing values were observed and were replaced with 'nan'. Also, categorical features with only two categories were encoded using mapping technique and features with more than two categories were encoded using get_dummies technique.

After preprocessing the dataset, the next step is to select the relevant features for the model. And some of the features which were not relevant for the model were dropped. The next step is to impute the missing values in the dataset. Since the values in the dataset were categorical, they were first encoded and then imputed using KNNImputer technique. After data transformation, the next step is to balance the dataset, as the dataset was highly imbalanced. So, RandomOverSampler technique was used to balance the dataset. Once the dataset is balanced, the next step is to build the model.

2.2. Model Implementation:

In this project, a classification model was built to predict the type of Thyroid a person has, based on the given features. We used LogisticRegression,KNeighborsClassifier,LinearSVC,SVC,MLPClassifier,RandomForestClassifier,GradientBoostingClassifierand BaggingClassifier to build models.

And We chose the RandomForestClassifier as it gave the highest accuracy.



```
In [70]: #checking accuracy scores of all models
         for name, model in models.items():
             print(name + ": {:.2f}%".format(model.score(X_test, y_test) * 100))
```

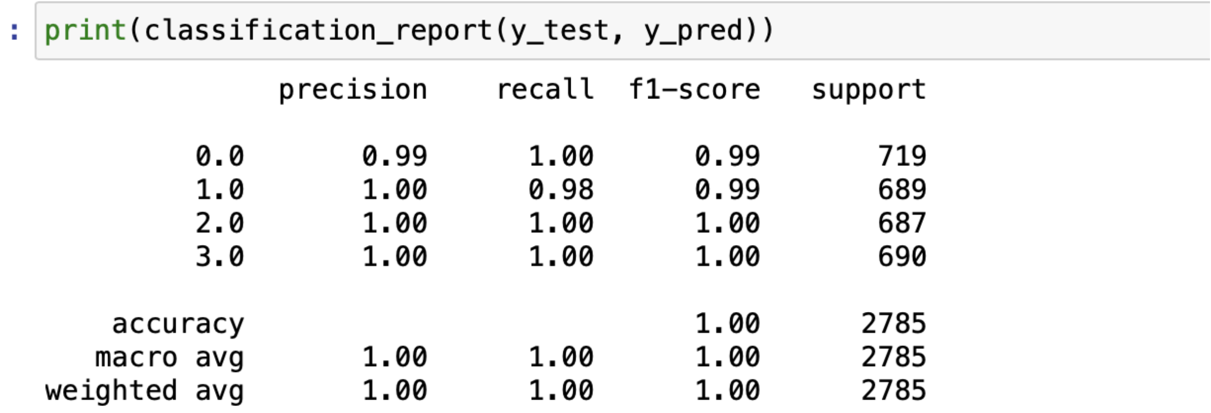
```
                   Logistic Regression: 77.99%
                   K-Nearest Neighbors: 96.98%
Support Vector Machine (Linear Kernel): 76.66%
   Support Vector Machine (RBF Kernel): 80.68%
                        Neural Network: 88.94%
                         Random Forest: 100.00%
                     Gradient Boosting: 92.21%
                    Bagging Classifier: 99.86%
```

The next step was to tune the hyperparameters of the RandomForestClassifier for better accuracy. We used RandomizedSearchCV to search for the best hyperparameters and assigned them to the model. To perform cross validation to check the accuracy of the model we used KFold and cross_val_score.

2.3. Model Evaluation:

To predict using the model, we used the predict method on the testing data to predict the output. We then plotted the confusion matrix using the sklearn library to view true-positive, false-positive, false-negative and true-negative.

Finally, we evaluated the model using accuracy score, cross-validation, confusion matrix and classification report.



```
: print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

         0.0       0.99      1.00      0.99       719
         1.0       1.00      0.98      0.99       689
         2.0       1.00      1.00      1.00       687
         3.0       1.00      1.00      1.00       690

    accuracy                           1.00      2785
   macro avg       1.00      1.00      1.00      2785
weighted avg       1.00      1.00      1.00      2785
```

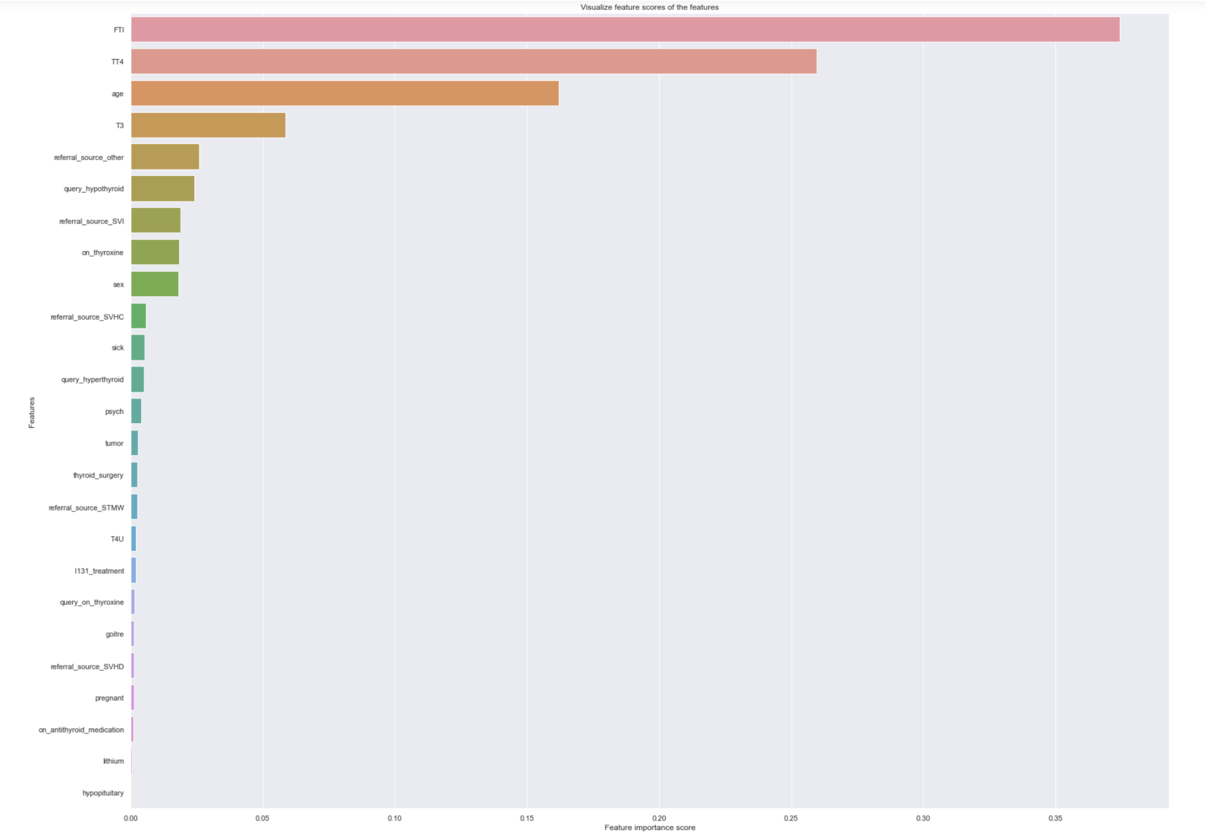
3. Conclusion

Our research examines the categorization of thyroid illnesses between hypothyroidism forms, due to the medical data that reveals major discrepancies in thyroid diseases. As the number of thyroid disease cases keeps rising globally, this is an important topic that needs to be addressed. By using machine learning algorithms, we were able to accurately classify this disease. The results of the machine learning application demonstrated a high level of accuracy. and the result of the accuracy of the random forest algorithm was 99.5% The Random Forest Classifier model has been used to classify the thyroid disease dataset with an accuracy of 99.5%, which is the highest accuracy among the other algorithms and a precision of 1.00. The model was tuned using hyperparameter tuning to get the best hyperparameter values. Cross fold validation was used to validate the model. The model was able to classify the thyroid disease dataset accurately and effectively. The confusion matrix was also plotted to visualize the results. The model was also able to predict the class of an unseen data point with an accuracy of 100%.

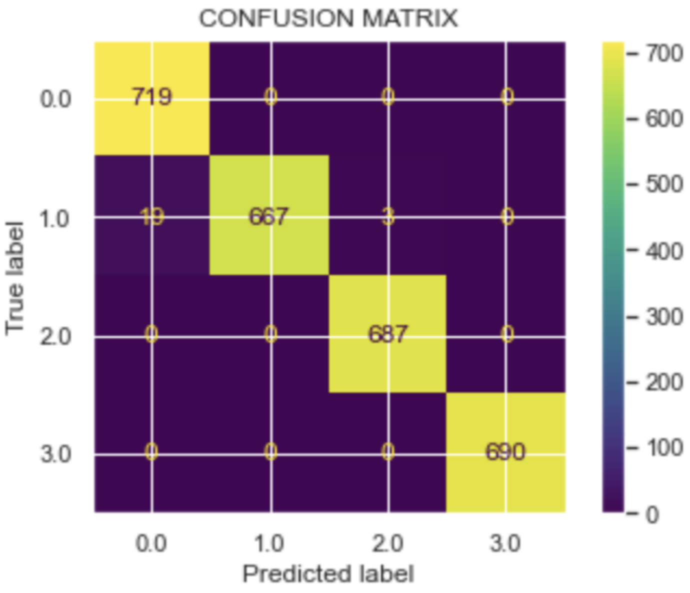
Overall, the machine learning based thyroid disease classifier has achieved a high accuracy and is able to correctly classify the data points into their respective classes. This model can be used to accurately diagnose the thyroid diseases in patients and can be used by medical professionals to provide better care.

4. Results

Feature importance graph assigns a score between 0 and 100 to each feature in a dataset, indicating its relevance in predicting the target variable.



From the above figure, we can see that there are numerous features which are not suitable for model training.



A confusion matrix is a tool used to evaluate the accuracy of a classification model. It consists of a table showing the model's predicted values compared to the actual values in the test dataset. The matrix helps to visualize the model performance by interpreting the number of correct and incorrect predictions made by the model. It also provides insight into which classes the model is predicting accurately and which classes it is misclassifying. The confusion matrix can be used to help identify areas of improvement for the model and can provide a good indication of its overall performance.

From the above confusion matrix we can conclude that there are more true positives and some false positive values. So, the overall model prediction is accurate enough to use the model.

5. Future Work

Future work on thyroid detection should focus on the development of more accurate, automated methods of diagnosing thyroid disease. By integrating the development of computer vision algorithms that can detect abnormalities in thyroid ultrasound images, or the development of machine learning models that can accurately predict thyroid disease from patient data.

Additionally, further research should be conducted into the causes of thyroid disease, as well as the development of better treatments. This research could lead to a better understanding of how thyroid diseases manifest, and the development of more targeted treatments that could reduce the risk of complications.

In the near term, research should focus on improving existing methods of thyroid diagnosis. This could involve improving the accuracy of radiographic and laboratory tests, as well as developing new biomarkers that could be used to detect thyroid diseases in their early stages. Finally, research should also focus on developing better methods of educating patients about their thyroid health and the importance of early detection and treatment.

Scope of improvement

- Improve accuracy of predictions by implementing more sophisticated and advanced machine learning algorithms such as Neural Networks and development of computer vision algorithms that can detect abnormalities in thyroid ultrasound images, or the development of machine learning models that can accurately predict thyroid disease from patient data.
- Incorporate additional data sources such as patient medical history, environmental factors, and characteristics of the thyroid and its surrounding tissue to form a more complete picture of the thyroid condition. Integrate more pre-processing steps, such as feature selection and dimensionality reduction, to help the machine learning models focus on the most important features in the data.
- Develop methods to detect and correct potential mislabeled data and bias in the data. And by focusing on improving the performance of the model when dealing with imbalanced data.

References

[1]. Chaubey, G.; Bisen, D.; Arjaria, S.; Yadav, V. Thyroid disease prediction using machine learning approaches. Natl. Acad. Sci. Lett. 2021, 44, 233–238. [CrossRef]

[2]. M. G. Ali and M. M. Al-Azab, “Computer-aided thyroid nodule classification using machine learning techniques,” PLoS One, vol. 13, no. 9, pp. 1–19, Sep. 2018.

[3] Dr. Srinivasan B, Pavya K “Diagnosis of Thyroid Disease: A Study” International Research Journal of Engineering and Technology Volume: 03 Issue: 11 | Nov – 2016

[4]. Ioni¸tˇa, I.; Ioni¸tˇa, L. Prediction of thyroid disease using data mining techniques. BRAIN Broad Res. Artif. Intell. Neurosci. 2016, 7, 115–124. 3. Webster, A.; Wyatt, S. Health, Technology and Society; Springer: Berlin/Heidelberg, Germany, 2020.

[5]. L. Li, J. Li, Y. Li, L. Li, and X. Zhang, “Multi-modal deep learning model for thyroid disease diagnosis,” in 2019 IEEE 5th International Conference on Computer and Communications (ICCC), 2019, pp. 1020–1024.

[6]. K. Liu, H. Chen, H. Wang, W. Fan, and M. Zhang, “Machine learning for thyroid nodule classification using ultrasound images,” Neurocomputing, vol. 320, pp. 103–113, Nov. 2018.

[7]. S. K. Gupta and S. B. Singh, “Thyroid disease detection system using machine learning,” in 2017 International Conference on Computing, Communication, and Automation (ICCCA), 2017, pp. 1184–1189.

[8]. J. Zheng, H. Yang, S. Huang, X. Yang and J. Zhang, “A novel convolutional neural network for thyroid nodule classification,” in 2019 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE), 2019, pp. 1–6.

[9]. N. Anuradha and C. V. Jawahar, “Thyroid disease detection using machine learning,” in 2016 International Conference on Computing Methodologies and Communication (ICCMC), 2016, pp. 467–471.

[10] Khushboo Taneja, Parveen Sehgal, Prerana “Predictive Data Mining for Diagnosis of Thyroid Disease using Neural Network” International Journal of Research in Management, Science & Technology (E-ISSN: 2321- 3264) Vol. 3, No. 2, April 2016