IJITCE

ISSN 2347- 3657

International Journal of
Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of
Information Technology & Computer Engineering

ISSN 2347–3657
Volume 12,Issue 2,April 2024

ThyroDectect: A Machine Learning Approach forThyroid Disease Prediction

T. Rajesh [1] P. Muni Sriya [2], Ambati Aarthi [3], Gurram Sukanya Devi [4],
P. Alekhya [5],

Abstract

According to the World Health Organization (WHO), approximately 300 million peoplesuffer from thyroid related conditions, with women being more prone to than men. Thyroid disorders, encompassing conditions like hyperthyroidism, hypothyroidism and thyroid nodule affect millions of people globally, with a significant impact on their health and well- being. The proposed system aims to revolutionize thyroid disorder prediction by leveraging machine learning algorithms. It utilizes a diverse dataset comprising patient demographics, medical history, and thyroid-related parameters, training models to classify individuals into four categories: Hypothyroidism,hyperthyroidism, negative, and sick. The objectives of the proposed system include improved enhance diagnostic accuracy, early detection, offer personalized predictions,and reduced healthcare cost.

Keywords: Thyroid detection, Thyroid Stimulating Hormones, Feature Selection, Class Imbalance, undersampling, oversampling, Hyperthyroidism, Hypothyroidism.

1. Introduction

Thyroid is a gland that is located in front of the neck and surrounds the windpipe, which releases the hormones such as (Tri-iodothyronine)T3, (thyroxine)T4 and Thyroid Stimulating Hormone (TSH). The thyroid primarily produces thyroid hormones, while the parathyroid glands regulate calcium and phosphorus levels through parathyroid hormone secretion. These glands are essential for maintaining proper metabolism, growth calcium balance. This glands also creates calcitonin, which facilitates the strengthening of bones. There are two conditions which are developed when thyroid gland does not function correctly they are hypothyroid a condition where the gland produces less amount of hormones whereas Hyperthyroidism is explored as another thyroid disorder characterized by excessive thyroid hormones production. Different types of hyperthyroidism, including Graves' disease, thyroid nodules, thyroiditis, and iodine excess, are explained.

[1:] Assistant Professor, Dept. of CSE (AI & ML), G.Narayanamma Institute ofTechnology and Science (For Women), India

[2, 3, 4,5:] Student, Dept. of CSE (AI & ML), G.Narayanamma Institute of Technologyand Science (For Women), India

The symptoms of hypothyroidism and hyperthyroidism depends on how serious the condition is, problem can take years to develop slowly. At initially, hypothyroidism symptoms like fatigue and weight gain could go unrecognised. Whereas as the symptoms of hyperthyroidism- are weight loss, tachycardia, diarrhea, anxiety, sweating, insomnia etc.

Dataset collected

Identifying and collecting relevant datasets containing information related to thyroid disorders, such as patient demographics, medical history, laboratory test results, and image data. Cleaned and pre-processed the available data by handling the missing values by replacing with median values and normalizing or scaling the features, at the end addressing any data quality issues. And finally splitting the data into training, testing and validation set. The dataset was collected from UCI repository, the entire dataset consists of 7200 instances and 22 attributes. It is classified into 3 different classes such as Normal (represented as 1), Hyperthyroidism (represented as 2), and Hypothyroidism (represented as 3).

2. Related Work

In this project, we worked on various Machine Learning Classifiers for reliably predicting thyroid disease, a major health concern that affects a considerable section of the population. Several authors concentrated on various Machine Learning techniques such as Naive Bayes, Support Vector Machine, Decision Tree, KNN, and Logistic Regression. S. Godara[4] demonstrated that Logistic Regression attained the best accuracy of 96.86% by employing various classification metrics such as accuracy, precision, recall, F1-score, ROC, and regression metrics such as Root Mean Squared Error. As shown by G. Chaubey[1], Decision trees attained the best accuracy of value 98.89% on the Irvin dataset of the UCI Repository.

Li-Na Li[2] showed that by utilizing 10-fold cross-validation, the best accuracy achievable for diagnosing thyroid illness is 97.73%. Using computer-aided diagnosis (CAD) which consists two important aspects known as rinciple Component Analysis (PCA) a dimensionality reduction technique and an Extreme Learning Machine (ELM) for rapid training.

Sagar Raisinghari conducted a comparison analysis, and the proposed system makes use of various ML algorithms to improve disease prediction accuracy. Decisiontree algorithm, which has an accuracy of 99.46%, is judged to be the best of them.

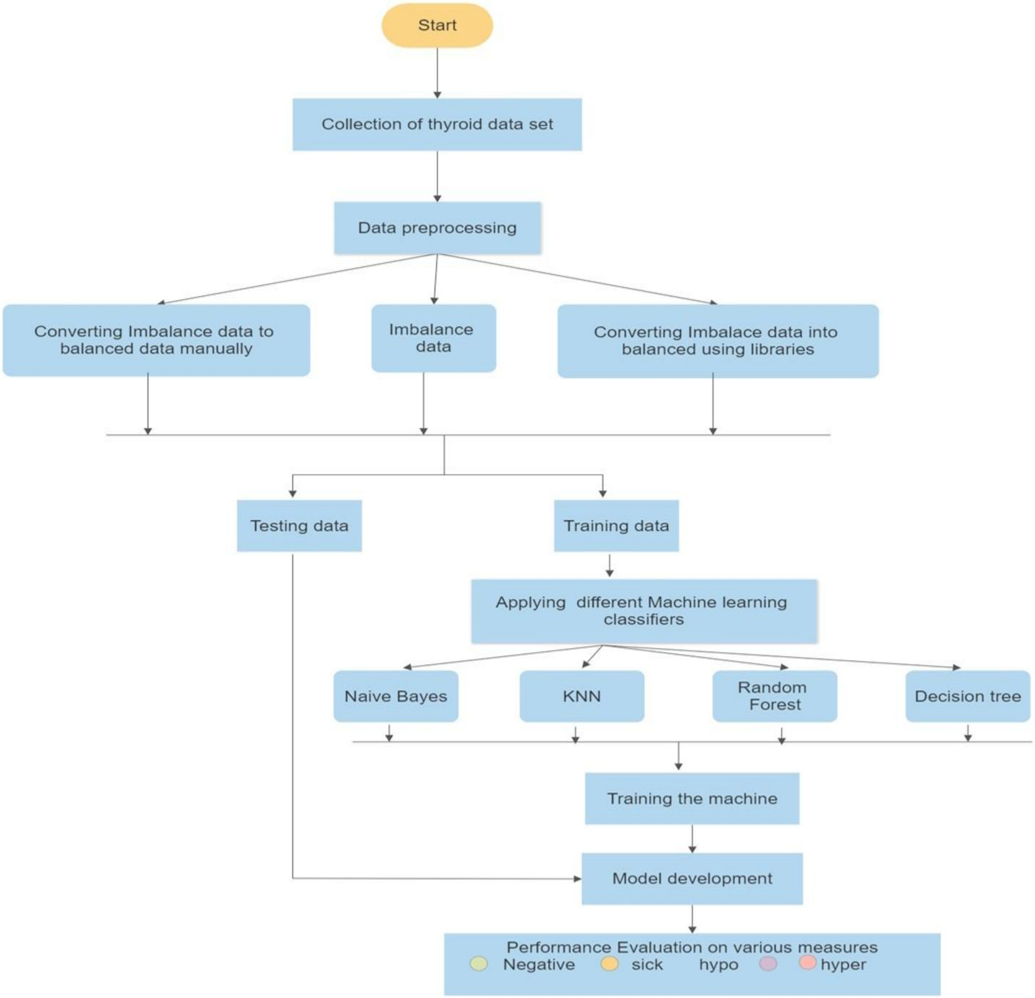
Lerina Aversano employed various machine learning techniques.

Particularly, we contrasted the output of 10 distinct classifiers. The various algorithms all display strong performances, particularly the Extra-Tree Classifier, whose accuracyexceeds 84%.

Dhyan Chandra Yadav and Saurabh Pal created the findings produced by individual classification methods such as decision tree which acquired an accuracy of98%, random forest tree achieved accuracy of 99%, and additional tree yield an accuracy 93%. Then, using the same dataset again, they created a bagging ensemble approach, which combined the three fundamental tree classifiers and provided agreater accuracy.

3. Proposed Methodology

3.1 Work flow


Start
Collection of thyroid data set
Data preprocessing
Converting Imbalance data to balanced data manually
Imbalance data
Converting Imbalace data into balanced using libraries
Testing data
Training data
Applying different Machine learning classifiers
Naive Bayes
KNN
Random Forest
Decision tree
Training the machine
Model development
Performance Evaluation on various measures
Negative
sick
hypo
hyper

3.2 Data pre-processing

The initial step in this process was to identify the important attributes that arenecessary for the predicting thyroid condition. Then comes the major step of eliminating missing values by identifying values such as “Na”, “NAN”, and “?” values, after identifying these values the missing values are replaced by the median values. We have also performed techniques like encoding such as one hot encoding forconverting the categorical or discrete data into the form on 1 or 0.

For identifying missing values:

```
for attributes in data.columns:
   value=data[attributes][data[attributes]=='?'].count() if value!= 0:
      print(attributes,data[attributes][data[attributes]=='?'].count())
```

Encoding of sex where:

sex = {ni: n for n, ni in enumerate(set(data['Sex']))}Female is consider as :0

Male is consider as :1

3.3 Feature Extraction

The feature extraction technique that is used is correlation analysis, statistical tests,and dimensionality reduction methods and balancing the imbalance data.

Correlation:

The Correlation matrix or a correlation graph is a way of visualizing andrepresenting the relationships between different attributes or variables in a dataset.

We have performed correlation technique in order to identify which attributesare effecting each other i.e. which attribute is dependent on each other and which are independent. We have done this analysis to eliminate the unnecessary attributes whichdo not have any effect on the result.

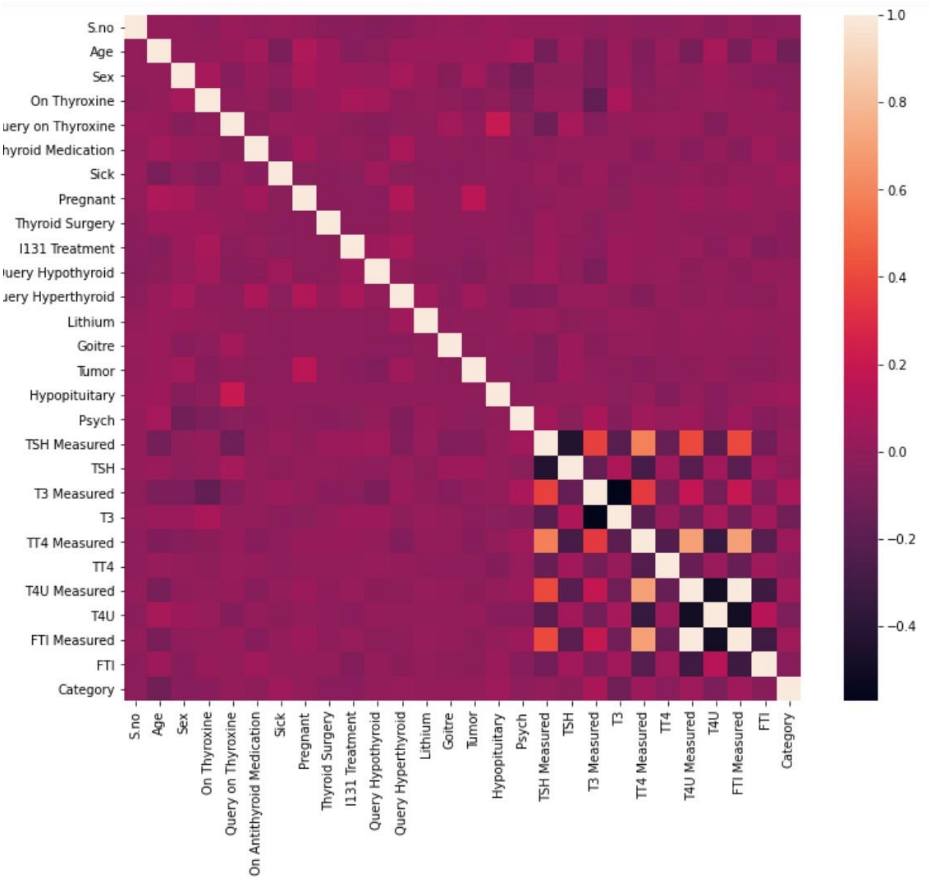


Fig:1 it describes about the correlation of different attributes and there behaviour.

Resampling the Data:

This Resampling techniques contain two sub categories they are Oversampling andUndersampling.

Oversampling occurs when there is an increase in the instances belonging to lower class.

Undersampling reduces the instances belonging to lowerclass.

According to the data that we have collected the data was imbalanced and bias

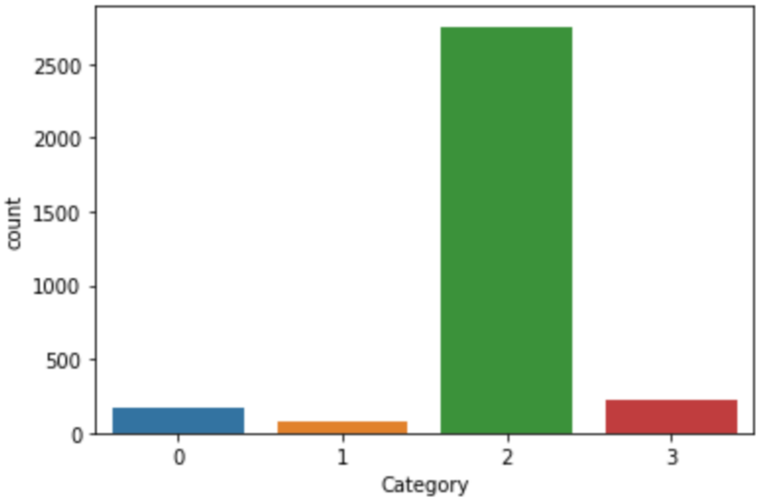


Fig:2 It shows the different class prediction before applying SMOTE analysis

The Fig:2 graph indicated that 0 :sick class, 1: hyperthyroid, 2:negative class, 3:hypothyroid as we can see that the negative class is oversampled so it leads to a biasdata prediction

In order to avoid that we have balanced that data by using the SMOT (Synthetic Minority Over Sampling Technique) technique. This technique is used to balance outthe data.

We are importing SMOTE technique from imblearn package.Before SMOTE:

Counter({2: 2200, 3: 179, 0: 134, 1: 63})

After SMOTE: Counter({2: 2200, 3: 2200, 0: 2200, 1: 2200})

As we can see from the above results that before SMOTE analysis the values are different, and after applying the SMOTE analysis all the class values are equated.

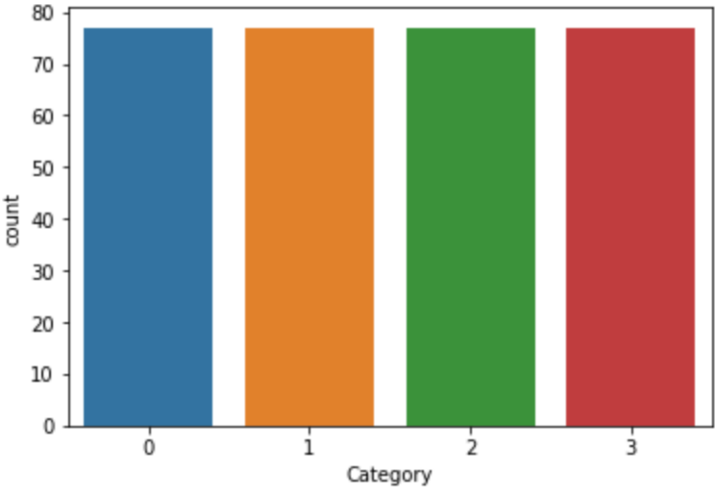


Fig:3 after applying SMOTE analysis and they are all correctly classifiedwithout any bias

3.4 Classification

In this phase, various machine learning classifiers are applied to train predictive models for thyroid condition. The following classifiers are employed : NaiveBayes, K-Nearest Neighbours (KNN), Decision Tree, and Random Forest.

Naïve Bayes

Naive Bayes is a fundamental classification method. Naïve Bayes classifier cannot keep up with classifiers such as decision trees, it occasionally surpass them in aparticular application areas, the most notable of which is text classification. The testingprocessing of the classifier becomes straightforward and affordable. Conditional probability refers to the likelihood of something happening if something else has previously occurred. Using conditional probability and previous information, we may determine the likelihood of an event.

$$\mathbf{P(H|E) = \frac{P(E|H) * P(H)}{P(E)}}$$

Where :

- P(H) is the probability that hypothesis H is right. The phrase for this is priorprobability. P(E) is the probability of the evidence (independent of the hypothesis).
- P(E|H) denotes the probability of the evidence if the hypothesis is right.
- P(H|E) represents the probability of the hypothesis if the evidence is available.

K-Nearest Neighbours (KNN)

The KNN method is supervised as well as non-parametric. The K-NN input is definedby the function space's K nearest occurrences. The usage of KNN for classification orregression has an effect on performance. The KNN method analyses the training data instances for the k-most related instances when undiscovered data instances require estimate. The prediction characteristics of the most comparable instances are aggregated and returned as the forecast for the undiscovered instance.

Decision Tree

Decision tree is a classifier which divides the entire dataset into different branches based on its entropy and information gain. Each and every internal node or non-leaf node represents a test on a specific property, each branch indicates the tests outcome,and leaf node a class name.

Random Forest

Random Forest is an ensemble learning approach that makes predictions by combiningnumerous decision trees. It employs the bagging principle and the randomization of feature selection to increase the model's generalization and decrease overfitting. The technique generates a number of decision trees, each trained on a randomly selectedpart of the training data and a selection of characteristics. To generate the final forecast, the predictions from separate trees are pooled via voting or averaging. The randomization of feature selection and sampling improves the resilience of the model.

We investigate several techniques to thyroid prediction by fitting these machine learning classifiers. Each algorithm has its own set of features and assumptions,allowing us to acquire a thorough grasp of its strengths and weaknesses in this context.

4. Results and Discussion

This project represents a comprehensive and meticulous exploration of thyroid prediction using the machine Learning techniques. The primary goal is to endeavour and to build a robust predictive model capable of accurately identifying thyroid conditions based patient data. These are the results that are opted after Appling different machine learning classifiers.

S.No	Algorithms	Accuracy
1	Naive bayes	45.1
2	K-Nearest Neighbor	50.6
3	Decision Tree	88.5
4	Random Forest	90.4

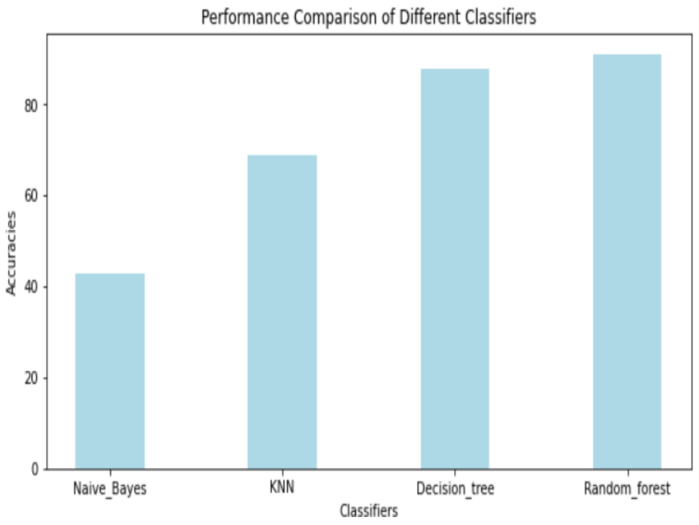


Fig:4

The above bar graphs show the different accuracies that are obtained from different machine learning classifiers, as we can see that Random forest is giving thehighest accuracy among all the 4 classifiers with an accuracy of 90.4 %.

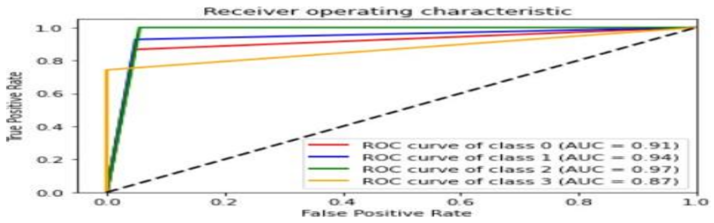


Fig:5

The fig: 5 is a ROC curve which indicates the evaluate the performance of a binary classification mode, such as random forest. The main reason of using this graph is it can easily distinguish between the positive class and the negative class across different threshold .

REFERENCES

1. Breiman, L. (2001). *Random Forests. Machine Learning*, 45(1), 5-32. ISSN: 0885-6125.
2. Chang, C. C., & Lin, C. J. (2011). LIBSVM: *A Library for Support Vector Machines*. ACM Transactions on Intelligent Systems and Technology, 2(3), 27:1-27:27. ISSN: 2157-6904.
3. Chen, Y., & Lin, H. (2018). *Thyroid Disease Classification Using Ensemble Machine Learning* Methods. Proceedings of the 2018 IEEE International Conference on Data Mining and Big Data, 167-172. ISBN: 978-1-5386-7694-2.
4. Chatterjee, S., & Nandi, S. (2019). *Comparison of Machine Learning Algorithms for Thyroid Disease Prediction*. International Journal of Computer Applications, 182(12), 10-15. ISSN: 0975-8887.
5. Gomes, T. M., & Plastino, A. (2020). *A Comprehensive Survey on Machine Learning Techniques for Thyroid Disease Prediction*. Journal of Medical Systems, 44(6), 120. ISSN: 0148-5598.
6. Kaur, A., & Kumar, V. (2016). *Data Preprocessing Techniques for Thyroid Disease Prediction*: A Comprehensive Review. International Journal of Computer Applications, 142(10), 6-11. ISSN: 0975-8887.
7. Ma, H., Wu, Y., & Peng, H. (2020). *Thyroid Disease Prediction Using Machine Learning Techniques*. International Journal of Computer Applications, 174(7), 1-5. ISSN: 0975-8887.
8. Park, S. J., An, J. H., Lim, J., & Oh, K. W. (2017). *Thyroid Disease Prediction by Considering Complex Relations among Clinical Factors*. Journal of Healthcare Engineering, 2017, 9187325. ISSN: 2040-2295.
9. Singh, P., Vishwakarma, D., Singh, V., & Malik, G. (2019). *Thyroid Disease Prediction Using Hybrid Machine Learning Techniques*. Proceedings of the 2019

International Conference on Artificial Intelligence and Computer Science, 25-30. ISBN: 978-1-4503-7245-6.

10. Zhang, H., Jiang, J., & Li, C. (2018*). A Comparative Study of Machine Learning Algorithms for Thyroid Disease Classification*. International Conference on Advanced Cloud and Big Data, 45-55. ISBN: 978-3-030-01093-6.

11. G. Chaubey, D. Bisen, S. Arjaria, V. Yadav, "Thyroid Disease Prediction Using Machine Learning Approaches," National Academy Science Letters, (June), 2020, doi:10.1007/s40009-020-00979-z.

12. L.N. Li, J.H. Ouyang, H.L. Chen, D.Y. Liu, "A computer aided diagnosis system for thyroid disease using extreme learning machine," Journal of Medical Systems, 36(5), 3327–3337, 2012, doi:10.1007/s10916-012-9825-

13. Y.I. Mir, S. Mittal, "Thyroid disease prediction using hybrid machine learning techniques: An effective framework," International Journal of Scientific and Technology Research, 9(2), 2868– 2874, 2020.

14. SunilaGodara, S.Kumar, "Prediction of Thyroid Disease Using Machine Learning Techniques," 10(2), 787–793, 2018.