Kurdish Studies
July 2024
Volume: 12, No: 5, pp. 466-470
ISSN: 2051-4883 (Print) | ISSN 2051-4891 (Online)
www.KurdishStudies.net

DOI: 10.53555/ks.v12i5.3247

Machine Learning Algorithms For Prediction Of Thyroid Syndrome At Initial Stages In Females

Syed Kanza Mehak[1*], Zeeshan Rasheed[2], Naeem Ahmed Ibupoto[3], Dr Shahzad Ashraf [4]

[1*]Hamdard University Karachi, Email: kanzamehak@gmail.com
[2]Department of Computer Science, MCKRU Sibi, Email: zeeshanrasheed1992@yahoo.com
[3]Department of Computer Science, GC University Hyderabad, Email: naeemloher@gmail.com
[4]DHA Suffa University Karachi, Email: nfc.iet@hotmail.com

***Corresponding Author:** Syed Kanza Mehak
*Email: kanzamehak@gmail.com

Abstract—In the modern world Machine learning plays a key role in the field of medical science, particularly in diagnosing health conditions and providing appropriate treatment at early stages. In the case of thyroid disease, traditional diagnostic methods involve detailed inspections and various blood tests. The primary aim and objective is to detect the syndrome of thyroid at initial stages with high level of accuracy. Machine Learning (ML) methods significantly enhance medical decision-making, accurate diagnosis, and reduce patient costs and time. This study aims to predict thyroid disease using dissimilar machine learning models, employing Random Forest, Support Vector Machine (SVM) and Logistic Regression algorithms. Thyroid patient dataset with relevant attributes is used to test these algorithms' effectiveness in diagnosing the disease. The outcomes determine the prospective of these machine learning methods in improving early diagnosis and treatment outcomes for thyroid patients.

Keywords—Machine learning techniques, ML, Thyroid disease, classification, SVM, Logistic Regression, Random Forest

INTRODUCTION

1.1 Machine Learning

Machine learning is a subfield of AI which emphases on emerging algorithms and statistical models that empower workstations to do certain jobs without being explicitly instructed. ML systems can progress their capacity over time by learning from and generating data-driven predictions. This capability has significant applications in a variety of fields, including healthcare, finance, marketing, and others. The fundamental premise of ML is to utilize information to train models that can recognize patterns, make choices, and provide insights. These models are built using a variety of methods, including supervised, unsupervised, and reinforcement learning.

1.2 THYROID

Thyroid gland is like a butterfly-shaped and positioned near the base of the human neck. The thyroid gland is critical to the preservation and stability of human metabolism, as well as the evolution and progress of the body. The thyroid gland primary tasks include blood circulation, temperature of body supervision, muscular strength, and brain function [1]. Somewhat injury or dysfunction of the gland might have serious ramifications for the proper functioning of the human body [2].Thus, proper thyroid hormone production results in a healthy human body. Low or excessive hormone levels will have a detrimental influence on human health. The thyroid gland generates numerous essential hormones, most notably (triiodothyronine), (thyroxine), and (thyroid-stimulating hormone) which play critical roles in regulating the body's metabolism

1.2.1 T3

T3 is one of the principal hormones generated by the thyroid gland, although it is also produced by the conversion of T4 in other tissues of the body. It is the most active of the two thyroid hormones and influences several physiological processes, including metabolism, heart rate, and body temperature. T3 enters cells and interacts with their DNA to control the metabolism of proteins, lipids, and carbohydrates.

1.2.2 T4

T4 (Thyroxine): The thyroid gland produces the most abundant hormone, T4. While T4 is not as active as T3, it functions as a precursor that is turned into T3 in the liver and other tissues. T4 regulates the muscle control, bone health, digestive processes, metabolism, brain development, and cardiac.

1.2.3 TSH

TSH is generated by the pituitary gland in the human brain rather than the thyroid gland. It stimulates the thyroid gland, hence regulating T3 and T4 synthesis. When T3 and T4 levels are low, the pituitary gland produces more TSH, which stimulates the thyroid to create additional hormones and when the same T3 and T4 levels are high, the pituitary gland inhibits TSH synthesis, which reduces thyroid hormone production. T3, T4, and TSH work together to maintain a delicate balance required for healthy metabolism and general body processes. Imbalances in these hormones can cause thyroid illnesses such as high and low level of hypothyroidism, each with its own set of symptoms and health concerns. The gland principally produces triiodothyronine (T3), thyroxine (T4), and thyroid stimulating hormone (TSH). Thyroid stimulating hormone (TSH) [3] is generated by the pituitary gland and primarily stimulates the thyroid gland to release T3 and T4, which boosts the metabolism of nearly every tissue in the body. As a result, the pituitary gland plays an important role in controlling thyroid hormone production as necessary. If TSH output is lower than T3, T4 secretion increases, and vice versa [3]. Thyroid disease is the most communal endocrine ailment in the globe. Thyroid disease varies from other endocrine illnesses in terms of treatment approach, relative attainability, and disease prediction [4]. Hypothyroidism is caused by inadequate thyroid hormone

secretion, and hyperthyroidism is caused by excessive secretion. Both conditions have a negative impact on human physiology, and indications of hyperthyroidism include dry skin, hair thinning, weight loss, high blood pressure, neck enlargement, and short menstrual cycles [1]. The symptoms includes thyroid gland swelling, rapid increase in weight, low blood pressure, heavy menstruation periods, and loss of craving. These symptoms may worsen if not treated immediately. As a result, an effective prediction model is necessary to help in the early detection of a patient's condition [9].

2. Literature Review

Bibi Amina Begum et al. [1] proposed many thyroid forecast algorithms using data mining, with an emphasis on T3, T4, and TSH, and studied classification methods such as Decision Trees, Backpropagation Neural Networks, SVM, and density-based clustering. Ankita Tyagi et al. [2] tested machine learning approaches including Decision Trees, SVMs, and K-nearest neighbors on a dataset from the UCI Machine Learning repository. Aswathi A K et al. [3] created a model with 21 characteristics and used partial swarm optimization to enhance SVM parameters. M. Deepika et al. [4] compared the accuracy of SVM, Decision Trees, and ANN across a wide range of illnesses, including thyroid problems. Sumathi A et al. [5] employed a decision tree technique to preprocess thyroid data before using the J48 classification. The thyroid data was preprocessed using a decision tree technique before being identified and created using the J48 classification algorithm. I Md. Dendi Maysanjaya et al. [6] tested several techniques and discovered that Multilayer Perceptron achieved the maximum accurateness of 96.74%. Ammulu K et al. [7] used the Random Forest approach and Weka to predict thyroid diseases from 25 factors. Roshan Banu et al. [8] studied data mining methods including Linear Discriminant Analysis and Decision Trees, while Dr. B. Srinivasan et al. [9] examined Decision Trees, Naïve Bayes, and SVM for thyroid diagnosis. Finally, Sunila Godara et al. [10] discovered Logistic Regression and Support Vector
Machines as promising models for thyroid prediction.

3. RESEARCH METHODOLOGY

The thyroid dataset is obtained from the Data Repository site [13]. The databank primarily contain records of female thyroid patient, which include all of the relevant information. In addition, the suggested model considers. Our proposed research is shown in below fig 1

3.1 Data Collection

The Thyroid Dataset, collected from the Kaggle Data Repository [13], provides critical information about female patients, including age, gender, pregnancy status, and any past health history. Particulars are stored in a databank and utilize in later clinical examinations. The dataset emphasizes traits that are highly linked to thyroid disease while ignoring less important ones. The characteristics might be true or false, or continuous values. The dataset's key characteristics include age, gender, hyperthyroidism, hypothyroidism, pregnancy, T3, T4, and TSH levels. The below Table I shows the attribute and type of attribute.

TABLE 1. DATASET DESCRIPTION

S.NO.	Attribute Name	Value Type
1	Age	1 to 90
2	Gender	Female
3	HyperThyroid	False, True
4	Thyroid	False, True
5	Pregnant	False, True
6	T3 value	Float value
7	T4 value	Float value
8	TSH value	Float value

The patient's past clinical history is also considered with collected Dataset, resulting in more trust worthy results. It primarily helps health care staffs to conduct complete investigation of the patient's condition which clearly depicts above in fig.1

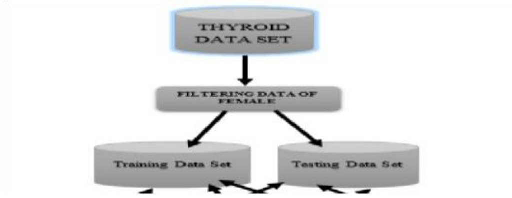


Fig 1 Proposed Research Methodology

In our study, we focus on analyzing thyroid data from female patients aged 1 to 90 years. To conduct our analysis, we first filter the dataset to include only the data of female patients within the specified age range. After filtering, we fragmented the dataset into training and testing datasets to build and evaluate our predictive models. The training set, which comprises 80% of the data, is used to train the model, while the remaining 20% is set aside as the testing set to validate the model's performance. This method ensures that our model can generalize well to new, unseen data, providing a robust basis for making accurate predictions about thyroid conditions in female patients. By using this split, we aim to develop a reliable tool for early detection and management of thyroid diseases, contributing to better health outcomes for women.

4. RESULTS AND DISCUSSION

In our study, we emphasis on analyzing thyroid data from female patients aged 1 to 90 years. This specific focus is chosen to ensure the relevance and accuracy of our findings, as thyroid conditions can vary significantly based on gender and age.

4.1 MACHINE LEARNING ALOGIRTHIM

Machine learning systems anticipate class labels based on example input data [12]. To create an efficient classification model, a training dataset is required, which helps the computer to understand patterns and correctly forecast [12]. This study analyzes three key predictive ML model: the Random Forest classifier, Logistic Regression, and the Support Vector Machine. It is well-known for its resistance to overfitting, especially on large datasets, as well as its ability to handle both numerical and categorical features. Random Forests are extremely beneficial for applications like as image and text classification [16]. Logistic Regression, on the other hand, is a statistical model developed specifically for binary classification problems. It employs a logistic function to express the likelihood of a class based on input features, making it simple and understandable. Logistic regression is commonly used in business and healthcare to forecast outcomes such as financial defaults or illness prevalence [16]. The Support Vector Machine is complex classifier that can handle both linear and nonlinear data by determining the optimum hyperplane to break groups. SVMs can manage complex linkages in data using a variety of kernel functions, making them appropriate for high-dimensional applications like image recognition and bioinformatics [17]

4.2 Random Forest

The Random Forest classifier achieved an accuracy of approximately 96.77%. The classification report and confusion matrix indicate that the model performs well, with high precision and recall for the negative class. However, the performance for the positive class (True) is lower, with a precision of 0.87 and recall of 0.69. The confusion matrix shows that there are some false positives and false negatives, but overall, the model is quite effective.

4.3 Logistic Regression:

The Logistic Regression model reached precision of about 90.62%. This means that it correctly predicted the class of data points nearly 91% of the time. The model shows a good balance in predicting both the positive and negative classes, with high precision and recall for each. However, it doesn't perform as well as the Random Forest model. Despite having some false positives (falsely evaluating the positive class) and false negatives (falsely evaluating the negative class), the Logistic Regression model remains a strong choice for classification tasks.

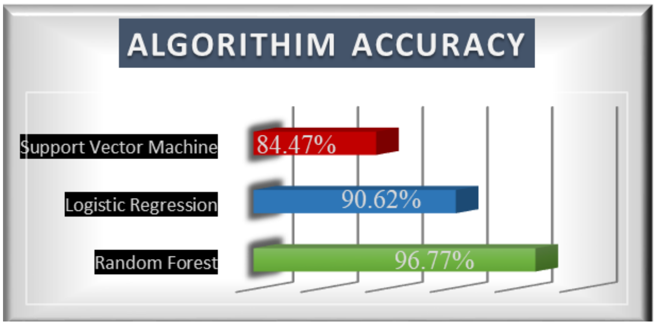
4.4 Support Vector Machine (SVM): The Support Vector Machine model reached precision of approximately 84.47%. This indicates that it correctly classified data points around 85% of the time. While it performs reasonably well, it doesn't match the accuracy of the Logistic Regression or Random Forest models. The SVM model tends to have more false positives and false negatives compared to the other two models, suggesting it has more room for improvement. Despite this, the SVM model still provides a reliable means of classifying data.
When comparing the Random Forest, Logistic Regression, and Support Vector Machine (SVM) classifiers, we can see distinct differences in their performance and characteristics

. TABLE 2 PERFORMANCE OF ALGORITHM

S.no	Algorithm	Accuracy	Performance
1.	Random Forest	96.77%	Best
2.	Logistic Regression	90.62%	Balanced

3.	Support Vector Machine	84.47%	Adequate

The Random Forest classifier stands out as the best performer with the highest accuracy and a wellbalanced precision and recall, especially for the negative class. This robustness makes it highly reliable for complex classification tasks where capturing true negatives is crucial. Logistic Regression offers a strong balance and high accuracy, making it a solid choice when a simpler and more interpretable model is preferred. It is particularly useful in scenarios where model transparency and ease of interpretation are important. The SVM, although slightly less accurate, remains a viable option, particularly if certain adjustments or improvements can enhance its performance. It surpasses in handling highdimensional data and can be tuned for better results. In summary, while all three models are effective, the Random Forest classifier is the most accurate and reliable, followed by Logistic Regression and then SVM. Each model has its own merits and demerits, but it depends on the study type of what specific requirement should be used, such as the need for model interpretability, computational efficiency, or the ability to handle complex data structures. Below graph clearly depicts that Random Forest algorithm best suit for initial stage detection of thyroid syndromes in female of dissimilar ages.



Graph 1 Algorithm Accuracy

5. CONCLUSION

Our study focuses on the analysis of thyroid syndrome data specifically in female patients aged 1 to 90 years, excluding data from male patients and irrelevant entries to ensure precision in our findings. By filtering and then split the data into training and testing sets, we aim to develop a robust predictive model. This model will help in the early detection and effective management of thyroid conditions among women, potentially leading to improved healthcare outcomes. Our targeted approach underscores the importance of demographic-specific research in enhancing the accuracy and applicability of medical predictions. By excluding data related to male patients, empty entries, and other irrelevant data, we can create a more homogeneous dataset that allows for more precise and reliable analysis. This targeted approach helps in understanding the patterns and factors affecting thyroid conditions among females, which can ultimately lead to more effective diagnosis and treatment strategies for this demographic.

On the other hand, our study of algorithms for thyroid syndrome in female patients reveals that the Random Forest is the most efficient classifier by achieving the highest accuracy of 96.77%. It demonstrates excellent performance, particularly in predicting. This makes it highly suitable for clinical settings where accurate diagnosis is critical. Overall, for the classification of thyroid syndrome in female patients, the Random Forest model is the preferred choice due to its superior accuracy and balanced performance, followed by Logistic Regression and SVM as a reliable and simpler option.

6. Future Work The study has mainly focused at initial stage diagnosis of thyroid syndrome in female. It has been suggested that this model can be extended by including more data not only females but also adding data of male.
(2021): 2581.

REFERENCES

1. Bibi Amina Begum and Dr. Parkavi "Prediction of thyroid Disease Using Data Mining Techniques" 5Th International Conference on Advanced Computing & Communication Systems (ICACCS), 2019
2. Ankith Tyagi, Ritika Mehra, Aditya Saxena "Interactive Thyroid Disease Prediction System Using Machine Learning Technique" 5th IEEE International Conference on Parallel, Distributed and Grid Computing(PDGC-2018), 20-22 Dec, 2018, Solan, India
3. Aswathi A K and Anil Antony "An Intelligent System for Thyroid Disease Classification and Diagnosis" Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE Xplore Compliant - Part Number: CFP18BAC-ART; ISBN:978-1-5386-1974-2
4. M Deepika and Dr. K. Kalaiselvi "A Empirical study on Disease Diagnosis using Data Mining Techniques." Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE Xplore Compliant - Part Number: CFP18BAC-ART; ISBN:978-1-5386-1974-2
5. Sumathi A, Nithya G and Meganathan S "Classification of Thyroid Disease using Data Mining Techniques" International Journal of Pure and Applied Mathematics, Volume 119 No. 12 2018, 13881-13890

6. Md. Dendi Maysanjaya, Hanung Adi Nugroho and Noor Akhmad Setiawan "A Comparison of Classification Methods on Diagnosis of Thyroid Diseases" 2015 International Seminar on Intelligent Technology and Its Applications
7. Ammulu K. and Venugopal T. "Thyroid Data Prediction using Data Classification Algorithm" IJIRST –International Journal for Innovative Research in Science & Technology| Volume 4 | Issue 2 | July 2017
8. Roshan Banu D and K.C.Sharmili "A Study of Data Mining Techniques to Detect Thyroid Disease" International Journal of Innovative Research in Science, Engineering and Technology (Vol. 6, Special Issue 11, September 2017)
9. Dr. Srinivasan B, K.Pavya "Diagnosis of Thyroid Disease Using Data Mining Techniques: A Study" International Research Journal of Engineering and Technology Volume: 03 Issue: 11 | Nov - 2016
10. SunilaGodara and Sanjeev Kumar "Prediction of Thyroid Disease Using Machine learning Techniques" International Journal of Electronics Engineering (ISSN: 0973-7383) Volume 10 • Issue 2 pp. 787-793 June 2018
11. A. Colubri, T. Silver, T. Fradet, K. Retzepi, B. Fry, P. Sabeti, Transformingclinicaldata into actionable prognosis models: machine learning Framework and fielddeployableapp to predict outcome of Ebola Patients, PLoSNegl. Trop. Dis. 10 (3) (2016) e0004549.
12. https://machinelearningmastery.com/types-of-classificationinmachine-learning
13. https://www.kaggle.com/kumar012/hypothyroid
14. Ali keles et al.,"ESTDD: Expert system for thyroid diseases diagnosis", Expert system with Applications,34,242-246,200
15. http://www.thehealthsite.com/diseasesconditions/world- thyroid - day- 2012 – facts-you-should-know/
16. Couronné, Raphael, Philipp Probst, and Anne-Laure Boulesteix. "Random forest versus logistic regression: a large-scale benchmark experiment." BMC bioinformatics 19 (2018): 1-14.
17. Zagajewski, Bogdan, et al. "Comparison of random forest, support vector machines, and neural networks for post-disaster forest species mapping of the krkonoše/karkonosze transboundary biosphere reserve." Remote Sensing 13.13