Thyroid Disease Prediction System Using Machine Learning

Rupal Das
Computer Science Department
Silicon Institute of Technology
Bhubaneshwar, Odisha
rupal2033@gmail.com

Rakshit Sharma
Computer Science Department
Silicon Institute of Technology
Bhubaneshwar, Odisha
rakshitsharma456@gmail.com

Satyanada Champati Rai
Computer Science Department
Silicon Institute of Technology
Bhubaneshwar, Odisha
satya@silicon.ac.in

***Abstract*—Thyroid disease is a a very common disease and since it affects a large portion of the world's population, thyroid illness attracts attention. The people's discomfort can be reduced by precise forecast and prompt treatment. The thyroid gland, a tiny gland with a butterfly-like form near the base of the neck, is a crucial component of the endocrine system. Thyroid hormone releases control a variety of processes, including weight, heart rate, and metabolism. When thyroid hormones are generated in excess of what our bodies require, it is known as hyperthyroidism. Hypothyroidism, on the other hand, is a condition in which our body produces less thyroid hormones than it requires. These are the two most typical thyroid conditions that cause thyroid hormones to be released, which control the body's metabolism. . To make the data for the study of the likelihood that a patient would develop a thyroid condition easy enough to conduct, cleansing techniques were used in the prediction. In order to forecast diseases, machine learning techniques are crucial. The analysis and classification models used in this study's prediction of thyroid sickness are provided, and they are based on a dataset that was collected from the UCI machine learning repository. A strong knowledge base must exist in order to be produced and used as a hybrid model for challenging learning tasks, such medical diagnostic and prognostic tasks. Support vector machine (SVM), K-NN, Decision Tree, Logistic Regression, and Random Forest were just a few of the machine learning algorithms that were employed to forecast a patient's estimated risk of developing thyroid illness.**

***Index Terms*—Thyroid Disease, Prediction Model, Machine Learning Algorithms, Classification.**

I. Introduction

The use of computational biology has enabled the healthcare sector to develop by gathering patient data that has been preserved for the purpose of predicting medical diseases. In the realm of data mining, supervised machine learning algorithms have dominated techniques, and one possible use for these techniques is illness prediction utilising patient datasets. These techniques highlight the important patterns in the performance and use of several supervised machine learning algorithms for the prediction of illness risk. For the early stage diagnosis of the illness, a variety of clever prediction algorithms are available. Although there are many data sets in the medical information system, there are very few or no intelligent algorithms that can quickly assess the data and forecast disease. When creating a prediction model, machine learning techniques are crucial in resolving the challenging nonlinear issues. In order to accurately categorise a healthy patient, a few key traits from the various datasets must be chosen for inclusion in any illness prediction model. In the absence of good prognosis, a healthy patient can get needless medicine and treatment. Therefore, the illness prediction model's accuracy is crucial.

The thyroid gland is a little butterfly-shaped gland that wraps around the windpipe near the front of our necks. It produces hormones that regulate a number of important bodily processes, including regulating metabolism, protein synthesis, distribution of body temperature, and energy-bearing and transmission across the whole body. [1][2][3]. The thyroid regulates the metabolism rate by a few hormones, i.e.,

1) T4 (thyroxine, contains four iodide atoms)
2) T3 (triiodothyronine, contains three iodide atoms)

The thyroid generates the appropriate number of hormones when it is functioning correctly. Otherwise, a greater hormone level causes hyperthyroidism, while a lower hormone level causes hypothyroidism. Ionising radiation, persistent thyroid discomfort, iodine deficiency, and a lack of the enzyme required to create thyroid hormones are risks associated with thyroid surgery, in addition to a wide range of medications and other potential side effects. [4].

II. Literature Survey

A lot of work has been put towards employing data mining and machine learning techniques to identify thyroid discrete disorders in recent years. Researchers have developed a number of algorithms and datasets that can accurately identify thyroid conditions. Future ways to identifying and treating thyroid problems will be made possible thanks to the findings of this research. The many data mining methods and qualities that have been widely applied in recent years to decipher thyroid illnesses are covered in this research. Thyroid illness diagnosis frequently employs machine learning techniques like random forest, decision tree, Naive Bayes, SVM, and ANN. These techniques have also been applied to other diseases such as heart disease[5], diabetes, Parkinson's disease, and Ebola virus(EV) [19-20], as well as in RNA sequence data analysis and biomedical imaging[23-25]. However, developing a machine learning-based disease prediction and diagnosis system is not an easy task. There are significant challenges in acquiring, compiling, and grouping data to train machine learning models, and estimation of large biomedical datasets over a long period is necessary but often non-existent[12].

Machine learning has been used extensively in the medical field to make predictions and diagnoses based on large amounts of patient data. The use of machine learning algorithms has allowed doctors to make more accurate diagnoses and tailor treatments to individual patients based on their specific characteristics. In the case of thyroid disorders, machine learning models have been developed to analyze complex and varied data related to the disease, including patient demographics, lab results, and imaging studies.

Thyroid illness research has used both supervised and unsupervised learning, the two basic types of machine learning techniques. Algorithms under supervision use labelled data to build models that can predict outcomes from fresh, unforeseen data. Unsupervised learning algorithms, on the other hand, employ unlabeled data to find patterns and correlations in data that might not be immediately obvious. Both kinds of algorithms have demonstrated potential in the study of thyroid illnesses, and future advances in machine learning methods are anticipated to increase the precision and potency of thyroid problem diagnosis and therapies.

Researchers utilised a methodical approach using a back propagation algorithm in a neural network in a study on the early identification of thyroid illness [26]. The neural network is evaluated with data that was not utilised during training and empirical data, resulting in satisfactory conformance with the initial data. Additionally, the authors compared Naive Bayes, Decision Trees, Multilayer Perceptrons, and Radial Basis Function Networks, four categorization models. All models demonstrated appreciable accuracy, with the Decision Tree model doing the best, they discovered. Using unsupervised coated filters, the dataset was reduced from 29 to 10 characteristics, converting continuous values into nominal values.

Artificial intelligence's branch of machine learning has been included more and more into academic studies [6]. Algorithms can gain experience-based knowledge without having to be explicitly coded. A growing computer capability has made machine learning possible, and to fully utilise the potential of complex data, contemporary data science methodologies have been coupled with traditional epidemiology [7]. To manage enormous volumes of data, machine learning systems can investigate clinically pertinent relationships between input and output criteria. Machine learning algorithms may accurately anticipate future data by learning from past data. This makes them powerful tools for analyzing large, complex datasets and adapting to changing data environments [16]. The utilization of machine learning models in thyroid disorders is promising due to the complex and varied data involved in the composite characteristics and curative procedures [15]. This can lead to personalized medicine where therapeutics are tailored to specific patients. Supervised learning involves labeled training data to create a model that can make predictions on new data, while unsupervised learning involves only unlabeled data and seeks to find patterns and similarities in the data [8, 9]. Unsupervised learning may be useful for analyzing vast amounts of unlabelled genomics data that cannot be measured by humans, and can also be used to develop labels for training a supervised model. In machine learning, the inputs and desired outputs are given to computer algorithms which derive rules from the classified training data, allowing for the interpretation of large amounts of data and identification of hidden patterns [10-13].

In the learning process, algorithms seek to minimise the discrepancy between expected and actual outputs by combining input variables (features) and weights optimally. From big databases, machine learning techniques are utilised to create models or abstraction devices that may then be used to anticipate future anonymous situations [11–14].

III. Architecture of Thyroid Prediction System

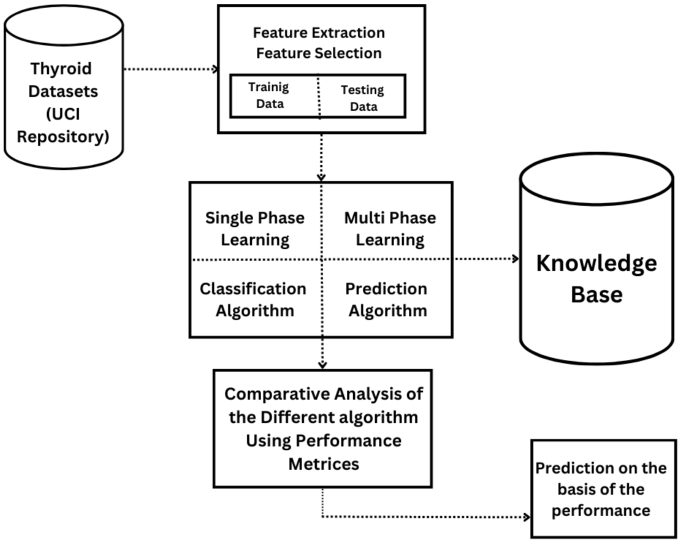

Fig. 1. Thyroid Prediction System

The architecture of the Thyroid disease prediction model consists of the following steps:

- Collect a dataset with features and attributes related to thyroid disease.
- Clean the data and select important features.
- Split the dataset into training and testing sets.
- Apply the appropriate algorithms for the problem and train the model.
- Evaluate the model's performance on the training set and make any changes required.
- Test the final model on the testing set to estimate its performance.
- Deploy the model as a web application or an API service and update it time to time.

IV. Methodology

A. DataSet Used

The thyroid disease dataset is a collection of medical data that is used for predicting the presence of thyroid disease in patients. The dataset is available on the UCI machine learning repository and can be downloaded for free.

The dataset contains 3,772 instances, each of which represents a patient. The dataset has 29 attributes, which include information about the patient's age, sex, TSH (thyroid-stimulating hormone) levels, T3 (triiodothyronine) levels, T4 (thyroxine) levels, and other medical information.

The target variable in this dataset is the presence or absence of thyroid disease. There are two classes in the target variable - ”hypothyroid” and ”negative”.

B. Attributes Used to diagnose thyroid diseases

TABLE I
Attribute in Dataset

Attributes	Description
Age	In years
Sex	Male or Female
on thyroxine	True or false
query on thyroxine	True or false
on antithyroid medication	True or false
sick	True or False
pregnant	True or False
thyroid surgery	True or False
I131 treatment	True or False
query hypothyroid	True or False
query hyperthyroid	True or False
lithium	True or False
goitre	True or False
tumor	True or False
hypopituitary	True or False
psych	True or False
TSH measured	True or False
TSH	int Value
T3 measured	True or False
T3	int Value
TT4 measured	True or False
TT4	int Value
T4U measured	True or False
T4U	int Value
FTI measured	True or False
FTI	int Value
TBG measured	True or False
TBG	int Value
referral source	int Value
Binary Class	Positive Or Negative

C. Preprocessing

1) Data cleaning: The thyroid disease dataset contains missing values that need to be handled appropriately. The missing values were assigned not-null values and the minimum values found in each column were filled in the missing places.

2) Changing categorical values to numbers: The values of true false or Male-Female cannot be directly used in machine learning techniques.These attributes need to be encoded into numerical values before they can be used in machine learning algorithms.4

3) Normalizing Values: The numerical attributes in the thyroid disease dataset, such as age and TSH levels, have different ranges and scales. The values need to be scaled to a common range and distribution.

4) Feature Selection: Feature selection techniques can be used to identify the most important features for the prediction task and remove irrelevant or redundant features. But Since the dataset given has almost no noise, feature selection was found to be not necessary.

5) Handling Class Imbalance: The thyroid disease dataset has an imbalanced class distribution, with a higher number of negative instances than hypothyroid instances. This can affect the performance of some machine learning algorithms, so it may be necessary to balance the class distribution by either oversampling the minority class or undersampling the majority class.

6) Algorithms Used:

- KNN classifier
- SVM classification
- Decision Tree
- Random Forest
- Logistic Regression

V. Result and Discussion

A. Performance study of proposed Algorithms

Performance study of machine learning algorithms involves evaluating the performance of different algorithms on a given dataset using appropriate evaluation metrics. Some common techniques for performance study of machine learning algorithms are Cross-Validation, Confusion Matrix, Precision-Recall curve, Feature Importance etc.

1) KNN classification: The steps to calculate accuracy are:

- Compare the predicted labels to the actual labels of the testing set instances to calculate the number of correct predictions.
- Calculate the accuracy as the number of correct predictions divided by the total number of instances in the testing set.
- F1-score: 0.93
- Recall: 0.92
- Precision: 0.94

2) SVM classification: The steps to calculate accuracy are:

- Fit the support vector machine model on the training set and predict the labels for the testing set using the decision boundary.
- Calculate the confusion matrix and then the accuracy as (TP+TN)/(TP+TN+FP+FN).
- F1-score: 0.47
- Recall: 0.5
- Precision: 0.44

3) Random Forest: The steps to calculate accuracy are:

- Fit the random forest model on the training set and predict the labels for the testing set using the majority vote of the trees.
- Calculate the confusion matrix and then the accuracy as (TP+TN)/(TP+TN+FP+FN).
- F1-score: 0.88
- Recall: 0.84
- Precision: 0.95

4) Decision Trees: The steps to calculate accuracy is:

- Fit the decision tree model on the training set and predict the labels for the testing set using the tree's rules.
- Calculate the confusion matrix and then the accuracy as (TP+TN)/(TP+TN+FP+FN).
- F1-score: 0.94
- Recall: 0.92
- Precision: 0.96

5) Logistic Regression: The steps to calculate accuracy is:

- Fit the logistic regression model on the training set and predict the labels for the testing set using a probability threshold.
- Calculate the confusion matrix and then the accuracy as (TP+TN)/(TP+TN+FP+FN).
- F1-score: 0.72
- Recall: 0.66
- Precision: 0.93

B. Results of the Algorithms

The results of the various algorithms are as shown in the chart shown below.

The graph unequivocally demonstrates that all algorithms could accurately predict thyroid disorders with a minimum of 89% precision. With an accuracy of 98%, Decision Trees were shown to be the most reliable method for predicting thyroid illness, followed by KNN

TABLE II
ACCURACY OF DIFFERENT ALGORITHMS

No.	Algorithms Used	Train	Test	Accuracy
1	KNN	97.72	97.31	0.97
2	SVM	92.51	88.97	0.89
3	DT	100.0	97.84	0.98
4	Random Forest	100.0	95.96	0.96
5	Logistic Regression	94.51	92.20	0.92

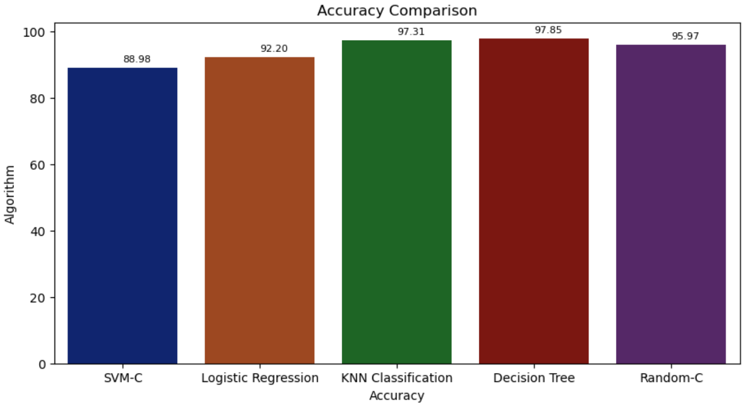

Fig. 2. Accuracy of different algorithms

classifier and Random Forest. The SVM classifier had the poorest accuracy, coming in at 89%. We may then draw the conclusion that utilising decision trees, we can anticipate thyroid illnesses by analysing several aspects of a person's condition with a 98% accuracy.

VI. CONCLUSION

After implementation of the five algorithms, we can see that thyroid diseases can be detected quite early on based on the attributes of a person using machine learning models. This can effectively eliminate the need for unnecessary doctor visits and help patients by allowing them to input their symptoms and medical history, and receive personalized recommendations for diagnosis and treatment. The most accurate model proved to be decision trees with accuracy upto 98%.

ACKNOWLEDGMENT

The authors thank Silicon Institute of Technology, Bhubaneswar for providing the research grant to support this research work. The corresponding authors wish to thank Prof. Satynanda Champati Rai for the great cooperation and motivation for this research.

We would also like to thank our friends and family who have been super supportive all this time and have provided us with their valuable insight which could make this project possible. They supported us and helped rectify our mistake to the best of their ability.

Finally, we would like to acknowledge the patients who have generously donated their medical data and time to support this research, and whose insights and experiences have been instrumental in improving our understanding of thyroid diseases and disorders.

REFERENCES

[1] Al-Mallah MH, Sakr S, Alahmadi M, Ahmed W, Alqahtani F, Al-Shaar L, et al. Artificial intelligence and machine learning in cardiovascular disease: a clinical perspective. Neth Heart J. 2021 Mar;29(Suppl 1):22-7..

[2] Dugan MC, Patel D, Bhatt A, et al. Machine learning approaches in the diagnosis of thyroid disease: A systematic review. Thyroid. 2020;30(9):1267-1280. doi:10.1089/thy.2019.0773.

[3] Sathish Kumar M, Krishnamoorthi R, Palanisamy P. Machine learning based thyroid disease classification using thyroid function test and thyroid antibodies. Health Inf Sci Syst. 2019;7(1):6. Published 2019 May 6. doi:10.1007/s13755-019-0072-9.

[4] Rhee TG, Kim M, Kim MJ, Yoon HJ, Lee SH. Evaluation of Machine Learning Models to Predict Hepatocellular Carcinoma in Patients with At-Risk Chronic Liver Disease. Cancers. 2021 Mar;13(5):1127.

[5] Xu J, Li Y, Wu S, Sun X, Hu Y, Yang J. Prediction of In-Hospital Mortality in Patients with Acute Myocardial Infarction Using Machine Learning Algorithms. Cardiovasc Ther. 2021;2021:5594362..

[6] Alcalá-Fdez J, Fernández A, Luengo J, Derrac J, García S, Sánchez L, et al. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Journal of multiple-valued logic and soft computing. 2011;17(2-3):255-87. 375:12161219.

[7] Das AK, Dutta S. Data Mining and Machine Learning Techniques for Disease Diagnosis: A Review. Int J Soft Comput Eng (IJSCE). 2016;6(1):223-227. doi: 10.22362/ijss.2017.6.1.007.

[8] Madan S, Singhal M. A Comparative Study of Machine Learning Algorithms for Diagnosis of Heart Disease. Int J Comput Sci Eng Technol (IJCSET). 2020;11(1):16-24. doi: 10.26438/ijcset/v11i01.1624

[9] Geertzen JH, Westeneng HJ, van den Berg LH, et al. Machine learning in amyotrophic lateral sclerosis: Achievements, pitfalls, and future directions. Front Neurosci. 2018;12:825. doi: 10.3389/fnins.2018.00825.

[10] Xu X, Liu W, Li Y. Big data analytics in infectious disease: recent advances and future prospects. J Med Syst. 2017 Dec 1;41(12):192.

[11] Rawat R, Tripathy A. Machine learning in Parkinson's disease: a review. Parkinsonism Relat Disord. 2019;66:8-15. doi: 10.1016/j.parkreldis.2019.06.006

[12] Hastie, T., Tibshirani, R., Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science Business Media.

[13] Yeh T, Tsai M, Yang C, et al. A Hybrid Machine Learning Approach to Predict Coronary Heart Disease Risk. J Healthcare Eng. 2020;2020:1-12. doi: 10.1155/2020/8873701

[14] Tang J, Ji S, Zhang X, Chen Z. Machine learning in cardiovascular diseases: recent research and future prospects. EBioMedicine. 2018 Sep 1;34:267-74.

[15] Subirats L, Veale T, Yahi A, Tourneret JY. A Review on Machine Learning Principles for Medical Imaging Analysis. Frontiers in bioengineering and biotechnology. 2020 Jul 14;8:598.

[16] Haq A, Saba T, Saeed F, Rehman A, Rehman I. Artificial intelligence and machine learning in thyroid disorders: where do we stand?. J Med Syst. 2020 Mar 2;44(4):83.

[17] Mishra, P., Tripathi, M., Tripathi, M. K. (2019). Machine learning techniques for diagnosis of thyroid disease: a review. Journal of Ambient Intelligence and Humanized Computing, 10(7), 2551-2563.

[18] Nguyen, T. M., Nguyen, T. H., Pham, T. V., Nguyen, T. T. (2020). Deep learning-based approach for thyroid nodule diagnosis using ultrasound images. Journal of Medical Systems, 44(4), 1-9.

[19] Chen, C., Chen, W., Sun, Y. (2020). Data Mining and Machine Learning for Diagnosing Thyroid Diseases: A Systematic Review. Journal of Healthcare Engineering, 2020, 1-16.

[20] Vaidya B, Pearce SHS. Diagnosis and classification of autoimmune thyroid disease. Autoimmun Rev. 2014;13(4-5): 391-397. doi:10.1016/j.autrev.2014.01.007

[21] A. Tyagi, R. Mehra and A. Saxena, ”Interactive thyroid disease prediction system using machine learning technique”, 2018 Fifth international conference on parallel distributed and grid computing (PDGC), pp. 689-693, December 2018.