IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 1, January 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 11, Issue 1, January 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1101007 |

# Detection and Classification of Hypothyroid Using Machine Learning

**Revati Rampur, Naveenkumar K Telkar, Nidhi Yashavanta, Avinash, Prof. Gnaneshwari G R**

Department of Computer Science and Engineering, Government Engineering College, Gangavathi, Karnataka, India

**ABSTRACT:** This paper discusses smart and precise ways to predict hypothyroid disease in a patient. Here the machine will be trained to detect whether the person has primary hypothyroidism or compensated hypothyroidism based on the user's input. So when a user enters data in a web app the data will be processed in the backend (model) and the result will be displayed on the screen. Thyroid related Blood test is used to detect the disease but it is often blurred and noise will be present. Data cleansing methods will be used to make the data primitive enough for the analytics to show the risk of patients getting the disease. Machine Learning plays a very deciding role in disease prediction. Machine Learning algorithms, KNN - K-Nearest Neighbor, random forest algorithms are used to predict the patient's risk of getting hypothyroid disease. Web app will be created to get data from users to predict the type of disease.

**KEYWORDS**: Hypothyroidism; KNN; Random Forest; K-Means; Flask

## I. INTRODUCTION

In an increasingly data-driven world, computational biology continues to be a fundamental means of developing emerging technologies for the field of evolutionary genomics. Computational biology refers to the creation and use of theoretical and data-analytical methodologies, computational simulation techniques, and mathematical modeling for the study of biological, behavioral, and social systems. This frequently means adopting a fresh perspective on biological systems, questioning accepted notions or ideas on the connections between system components, or integrating data from many sources to create a model that is more thorough than previous attempts. Machine learning has become a pivotal tool for many projects in computational biology, bioinformatics, and health informatics. Machine learning was able to grow quickly and become widely employed in the computational biology field due to its unique capacity to handle massive datasets and generate predictions on them using precise statistical models. So far, machine learning has been used to solve a variety of computational biology challenges, assisting researchers in learning more about a wide range of biological topics. Most Machine learning models perform well due to their custom-designed representation and input features. Using the input data generated through that process, ML learns algorithms, optimizes the weights of each feature, and optimizes the final prediction. With the help of prediction algorithms detection and classification for a disease can be made at early stages.

## II. LITERATURE SURVEY

In paper [1], authors used Machine Learning algorithms such as SVM - support vector machine, decision tree, logistic regression, KNN - K-nearest neighbor, ANN- Artificial Neural Network to predict the patient's risk of getting thyroid disease.Data cleansing methods were used to make the data primitive enough for the analytics to show the risk of patients gettingthis disease. Authors made use of logistic regression algorithm to train the dataset and to predict thyroid disease with more accuracy. Here the machine was trained to detect whether the person normal, hyper-hypothyroidism based on the user's input. Among which accuracy obtained from the logistic regression model (96.92%) was highest, this model was considered for developing the prediction model.The work offered good accuracy in terms of prediction of thyroid disease. The primitive webpage was able to predict thyroid disease based on individual patient data, but there is still a possibility for human error when entering the data and the model may take time to process multiple patient records.

According to Gomathy, C.K[6] Machine learning techniques can be used to help detect diseases by analyzing large amounts of data and identifying patterns or trends that may not be apparent to humans. Once trained, the algorithms can be used to analyze new data and make predictions about the likelihood of a person having a particular disease. Use of machine learning in disease detection is in the identification of risk factors for developing certain diseases. By analyzing data on factors such as age, lifestyle, and genetics, machine learning algorithms can help predict a person's

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 11, Issue 1, January 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1101007 |

risk of developing a particular disease and provide recommendations for reducing that risk. Machine learning techniques have the potential to significantly improve the accuracy and efficiency of disease detection, ultimately leading to better patient outcomes.

In paper [7], the authors examined the use of four different classification models - Naive Bayes, Decision Tree, Multilayer Perceptron (MLP) and Radial Basis Function (RBF) Network - on a thyroid data set in order to more accurately identify thyroid disease, specifically hyperthyroidism and hypothyroidism. The study used data mining techniques to conduct a comparative analysis of different algorithms and their effectiveness in predicting thyroid disease. This allowed for a more thorough examination of the different classification models, ultimately leading to the identification of the most effective model for predicting thyroid disorders.The study found that the accuracy of the classification models varied in different levels of experiments possibly due to a lack of information in the database and the quality of the data used.The decision tree model was found to be the most effective in all experiments.

Dahiwade et al. [8] proposed a ML based system that predicts common diseases. The symptoms dataset was imported from the UCI ML repository, where it contained symptoms of many common diseases. The system used CNN and KNN as classification techniques to achieve multiple diseases prediction. Moreover, the proposed solution was supplemented with more information that concerned the living habits of the tested patient, which proved to be helpful in understanding the level of risk attached to the predicted disease. Authors compared the results between KNN and CNN algorithm in terms of processing time and accuracy. The accuracy and processing time of CNN were 84.5% and 11.1 seconds, respectively. The statistics proved that KNN algorithm is underperforming compared to CNN algorithm.

| | | | | | |
|---|---|---|---|---|---|
| 1 | Chandan R et al. | Thyroid Detection Using Machine Learning | In this work they used Machine Learning algorithms such as SVM - support vector machine, decision tree, logistic regression, KNN - K-nearest neighbor, ANN-Artificial Neural Network to predict the patient's risk of getting thyroid disease. | The work offered good accuracy in terms of prediction of thyroid disease. | The primitive webpage was able to predict thyroid disease based on individual patient data, but there is still a possibility for human error when entering the data and the model may take time to process multiple patient records. |
| 2 | Lerina Aversanoet al. | Thyroid Disease Treatment prediction with machine learning approaches | Differently from other studies, that mainly aim to detect the disease, this work focused on the clinical history of patients suffering from hypothyroidism and treated with thyroid hormone to predict the treatment trend. | The model was able to predict the progress of the patient's treatment on the basis of other parameters related to the person being treated, therefore helping the doctor in choosing the dosage of the drug to prescribe. | One limitation of this study was the quality of the data set used. It did not take intoaccount the presence of any secondary thyroid diseases, which could affect the accuracy of the findings. |
| 3 | Yongfeng Wang and Wenwen Yue | Comparison Study of Radiomics and Deep Learning-Based Methods for Thyroid Nodules Classification Using Ultrasound Images | This study compared the performance of two different methods for classifying thyroid nodules based on ultrasound images: radiomics and deep learning. | This allowed them to directly compare the strengths and weaknesses of these two methods and determine which one offers better performance. | In the following study, the accuracy of the deep learning method only achieved 74.69%. In the deep learning based method, they only used a common CNN model (VGG16) as a pre-trained model and trained it in a |

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 11, Issue 1, January 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1101007 |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | relatively smaller dataset. |
| 4 | Borzouei S, Mahjub H et al. | Diagnosing thyroid disorders: Comparison of logistic regression and neural network models | The objective of this study was to investigate the potential of multinomial logistic regression and neural network models for diagnosing the two most common thyroid disorders, hyperthyroidism and hypothyroidism. | Results based on this study showed better performance of neural network model than multinomial logistic regression in all cases. | The limitation with this study was that this study showed less accuracy as regards to the logistic regression compared to our study which obtains higher accuracy. |
| 5 | Singh, Nikita and Jindal, Alka | A Survey of different Types ofCharacterization Technique in Ultra Sonograms of the Thyroid Nodules | The present study focused on all thyroid classification approaches and texture characterization approaches. This survey described different methods for thyroid feature extraction and thyroid classification used in medical images. | This study concluded that among all the methods discussed for thyroid classification, fuzzy local binary pattern (FLBP) provided higher classification performance compared to the radon-based approach and the GLCM approach. | Sometimes the LBP can miss the local structure as they don't consider the effect of the center pixel. The binary data produced by them are sensitive to noise. |

## III. PROPOSED SYSTEM

- We propose a model based on machine learning as a solution for detecting thyroid disease. The model will be trained on a large dataset of labeled medical data, or blood test results, to recognize patterns indicative of thyroid conditions.
- It is important to carefully validate and test the model before implementing it in clinical practice, to ensure its reliability and accuracy.

There are several potential advantages to the proposed system for detecting and classifying thyroid disease using a machine learning model:

- Accuracy: The model has the potential to be more accurate than human observers at detecting thyroid disease, particularly if it is trained on a large, diverse dataset of labeled medical data.
- Efficiency: The model can process data quickly and consistently, potentially reducing the time and resources required for thyroid detection.
- Consistency: The model can provide consistent, reliable results, as it is not subject to human bias or variability.
- Scalability: The model can be easily scaled to process large amounts of data, potentially making it more efficient and cost-effective for use in large healthcare systems.

## IV. CLASSIFICATION TECHNIQUES IN MACHINE LEARNING

### A. K-Means Clustering

K-Means Clustering is a well-known unsupervised machine learning algorithm that is commonly used to group similar data points together into clusters. It is an iterative algorithm that starts by randomly selecting a specified number of cluster centers, or "means," and then assigns each data point to the closest mean. The algorithm then adjusts the cluster means based on the data points assigned to them, and continues to iteratively reassign and adjust the data points and cluster means until the clusters reach convergence and become stable.

The goal of K-Means Clustering is to partition the data into clusters in a way that minimizes the total within-cluster sum of squares. This measure reflects the sum of the squared differences between the data points in a cluster and the

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 11, Issue 1, January 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1101007 |

cluster mean, and is used to quantify the tightness or cohesiveness of the clusters. By minimizing this measure, the algorithm aims to create clusters that are as compact and coherent as possible.

K-Means Clustering is often used for tasks such as feature learning, where it can be used to identify patterns and trends in the data, and to group similar data points together. It is also commonly used for exploratory data analysis and visualization, as it can help to reveal the underlying structure of the data and to identify any unusual or outlier observations.

B. K-Nearest Neighbour

K-Nearest Neighbour (K-NN) is a simple and widely used machine learning algorithm that is based on the supervised learning method. It is a type of instance-based learning, which means that it stores all of the training data and makes predictions based on the similarity of new data to the stored data.

To classify a new data point using the K-NN algorithm, we first need to specify the value of K, which determines the number of nearest neighbors that will be considered when making the prediction. The algorithm then compares the new data point to the K nearest neighbors in the training data, based on some measure of similarity or distance, and assigns the new data point to the category that is most common among the K nearest neighbors.

One of the main advantages of the K-NN algorithm is that it is easy to understand and implement, and it can be used for a wide range of classification tasks. It is also relatively fast and efficient, as it does not require any explicit training or model fitting, and it can make predictions based on the stored data in real-time. However, it can be sensitive to the choice of K and to the scale and distribution of the data, and it may not always perform well on complex or highly non-linear data.

C. Random Forest

Random Forest is a popular and widely used machine learning algorithm that is based on the supervised learning technique. It is a type of ensemble learning algorithm, which means that it combines the predictions of multiple classifiers in order to make more accurate and reliable predictions.

The Random Forest algorithm works by creating a large number of decision trees using different subsets of the training data, and then averaging the predictions made by each tree. Each tree is trained using a random subset of the data, and it makes predictions based on the features and patterns it has learned from that subset. By aggregating the predictions of multiple trees, the random forest algorithm is able to reduce the variance and overfitting that can occur when using a single decision tree, and to improve the overall predictive accuracy of the model.

Random Forest is a flexible and powerful algorithm that can be used for a wide range of classification and regression tasks. It is generally considered to be a robust and reliable algorithm, and it is often used as a baseline or benchmark for comparison with other machine learning algorithms. However, it can be computationally intensive to train and use, and it may not always perform well on very high-dimensional or sparse data.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 11, Issue 1, January 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1101007 |

## V. METHODOLOGY

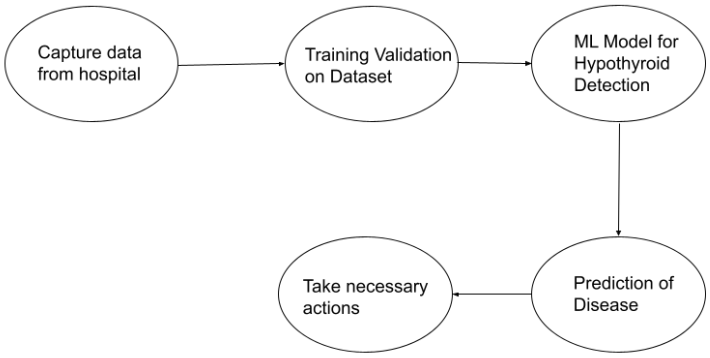


Fig1. Proposed Methodology

### A. Data Description

We will be using the Thyroid Disease Data Set present in the UCI Machine Learning Repository. It consists of patients' thyroid records. The records include different attributes, such as age, sex, and thyroid hormone levels. Total 7200 instances present in different batches of data. It has 29 columns in totalas given in Table 1.

| Sl. No. | Attribute Name | Value Type |
|---|---|---|
| 1 | age | continuous |
| 2 | sex | M,F |
| 3 | on_thyroxine | t,f |
| 4 | query_on_thyroxine | t,f |
| 5 | on_antithyroid_medication | T,f |
| 6 | Sick | T,f |
| 7 | Pregnant | T,f |
| 8 | thyroid_surgery | t,f |
| 9 | I131_treatment | t,f |
| 10 | query_hypothyroid | t,f |
| 11 | query_hyperthyroid | t,f |
| 12 | lithium | t,f |
| 13 | goitre | t,f |
| 14 | tumor | t,f |
| 15 | hypopituitary | t,f |
| 16 | psych | t,f |
| 17 | TSH_measured | t,f |
| 18 | TSH | continuous |
| 19 | T3_measured | t,f |
| 20 | T3 | continuous |
| 21 | TT4_measured | t,f |
| 22 | TT4 | continuous |
| 23 | T4U_measured | t,f |
| 24 | T4U | continuous |
| 25 | FTI_measured | t,f |
| 26 | FTI | continuous |
| 27 | TBG_measured | t,f |
| 28 | TBG | continuous |
| 29 | referral_source | continuous |

**Table 1. Data Descriptiom**

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 11, Issue 1, January 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1101007 |

B. Data Preprocessing

In this step we will export all batches of data from the database into a single csv file for training. Alsowe explore our data set here and do Exploratory Data Analysis also known as EDA if required and perform data preprocessing depending on the data set. We first explore our data set in Jupyter Notebook and decide what pre-processing and validation needs to be done such as imputation of null values, dropping some column, etc and then we have to write separate modules according to our analysis, so that we can implement that for training as well as prediction data.

C. Data Clustering

The K-Means algorithm will be used to create clusters in the pre-processed data. The optimum number of clusters is selected by plotting the elbow plot. The idea behind clustering is to implement different algorithms to train data in different clusters. The K-means model is trained over pre-processed data and the model is saved for further use in prediction.

D. Get best model of each

In this step, we will use the data clustering algorithm to divide the dataset into multiple clusters. Then, for each cluster, we will train several different machine learning models and evaluate their performance. Finally, we will select the best model for each cluster, based on its performance on the training data. This will allow us to obtain a set of models that are tailored to the specific characteristics of each cluster, and that can be used to make predictions or decisions about new data in that cluster.

E. Hyper parameter Tuning

After selecting the best model for each cluster, we will do hyper parameter tuning for each selected model. By doing hyperparameter tuning for each selected model, the goal is to improve the performance of these models on the specific clusters they were chosen for.

F. Model Saving

After we have fine-tuned the hyperparameters of our models through the process of hyperparameter tuning, we will save these optimized models for future use. Saving the models allows us to use them for prediction purposes without the need to retrain them, saving time and resources. This can be useful if we need to make predictions on new data or if we want to deploy the model in a production setting.

G. Cloud Setup

In this step of the process, we will prepare the cloud infrastructure for deploying our trained model. This may involve setting up a cloud computing platform, such as Amazon Web Services or Google Cloud Platform, and configuring the necessary resources and permissions. We will also create a web application using the Flask framework, which will provide a user interface for interacting with the model. Finally, we will integrate the trained model with the Flask app and UI, allowing users to input data and receive predictions from the model through the web interface.

H. Push app to cloud

Once we have set up the cloud environment and tested the web application locally, we will deploy the app to the cloud in order to make it available for use. This may involve uploading the code and necessary resources to the cloud platform and starting the application. After the app has been deployed to the cloud, it will be accessible to users through the internet.

I. Prediction

Now that our application on cloud is ready for doing prediction. The prediction data will be exported from DB and further will do the same data cleansing process as we have done for training data usingmoduleswe will write for training data. User data will also go along the same process of Exporting data from DB, Data pre-processing, Data clustering and according to each cluster number we will use our saved model for prediction on that cluster.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 11, Issue 1, January 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1101007 |

J. Export Prediction to CSV

After obtaining the predictions for the client's data, our final task will be to export the predictions to a CSV file. This will allow us to save the predictions in a format that is easy to view and share. Exporting the predictions to a CSV file will allow us to easily review and analyze the results of our machine learning model.

## VI. CONCLUSION

The Detection and Classification of Hypothyroid Disease using Machine Learning project is focused on using advanced techniques to accurately predict thyroid disease. To achieve this goal, we will employ a variety of machine learning algorithms and techniques, including K-Means Clustering, K-Nearest Neighbour, and Random Forest.

We will have used these algorithms to cluster the data and train our models, and we have selected the best performing model for making predictions on new, unseen data. To evaluate the performance of our model, in the future, we will use a range of evaluation metrics to determine the accuracy of our prediction model for the presence of hypothyroid disease."

In the future, our approach using machine learning techniques will have demonstrated promising results in the detection and classification of hypothyroid disease. We believe that it will have the potential to be a valuable tool in clinical settings. By leveraging the power of machine learning, we will be able to more accurately and reliably predict and classify thyroid disease.

## VII. FUTURE ENHANCEMENTS

As we move forward with this project, we envision expanding its scope to include the development of a machine learning model that can predict thyroid disease with even greater accuracy and efficiency. This model will be based on advanced algorithms and techniques that have been carefully chosen to optimize its performance. Additionally, by deploying the model in the cloud, we aim to make it easily accessible to a wide range of users. With this approach, we hope to provide a valuable tool that can help improve the diagnosis and treatment of thyroid disease for people around the world.

There are several advantages to this:

- Wide accessibility: By deploying the model in the cloud, it can be easily accessed by a wide range of users from any location with an internet connection. This can help to make the model more widely available to healthcare professionals and patients around the world.
- Cost-effective: Machine learning models can potentially reduce the costs associated with diagnosing thyroid disease by streamlining the diagnostic process and reducing the need for manual labor.
- Personalized predictions: Our models can analyze an individual's specific data, such as their medical history and test results, to make personalized predictions about their likelihood of having thyroid disease.

Batch Predictions: Our machine learning model's ability to perform batch predictions allows for efficient and accurate analysis of multiple data points, improving its utility and value for a variety of applications including streamlining the diagnostic process and large-scale epidemiological studies.

## REFERENCES

1. Chandan R & Chetan Vasan “Thyroid Detection Using Machine Learning”, International Journal of Engineering Applied Sciences and Technology, 2021 Vol. 5, Issue 9, ISSN No. 2455-2143, Pages 173-177. Published Online January 2021 in IJEAST (http://www.ijeast.com)
2. Lerina Aversano et al. “Thyroid Disease Treatment prediction with machine learning approaches”, Procedia Computer Science, Volume 192, 2021, Pages 1031-1040, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2021.08.106. (https://www.sciencedirect.com/science/article/pii/S1877050921015945)

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 11, Issue 1, January 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1101007 |

3. Y. Wang et al., "Comparison Study of Radiomics and Deep Learning-Based Methods for Thyroid Nodules Classification Using Ultrasound Images," in IEEE Access, vol. 8, pp. 52010-52017, 2020, doi: 10.1109/ACCESS.2020.2980290.
4. Borzouei S, Mahjub H, Sajadi NA, Farhadian M. "Diagnosing thyroid disorders: Comparison of logistic regression and neural network models". J Family Med Prim Care. 2020 Mar 26;9(3):1470-1476. doi: 10.4103/jfmpc.jfmpc_910_19. PMID: 32509635; PMCID: PMC7266255.
5. Singh, Nikita and Jindal, Alka (2012) "A Survey of Different Types of Characterization Technique In Ultra Sonograms of the Thyroid Nodules," International Journal of Computer Science and Informatics: Vol. 2: Iss. 2, Article 14. DOI: 10.47893/IJCSI.2012.1080 Available at: https://www.interscience.in/ijcsi/vol2/iss2/14
6. Gomathy, C K. (2021). The Prediction Of Disease Using Machine Learning.
7. S. Razia, P. SwathiPrathyusha, N. Krishna, and N. Sumana. "A Comparative study of machine learning algorithms on thyroid disease prediction", International journal of engineering and technology, 7:315, 2018
8. D. Dahiwade, G. Patle, and E. Meshram, "Designing disease prediction model using machine learning approach," Proceedings of the 3rd Inter- national Conference on Computing Methodologies and Communication, ICCMC 2019, no. Iccmc, pp. 1211–1215, 2019.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details