JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Thyroid disease prediction using Machine Learning

K.Keerthana[1] ,P.Murali Krishna[2] ,Ch.Sai Sri Harsha Vardhan [3],K. Saketh[4],MD.Abdul Rahman[5]

Assistant Professor [1] , Students[2][3][4][5]

Vignana Bharathi Institute Of Technology,Hyderabad,Telangana.

Abstract : We developed a new system to predict thyroid disease using two powerful machine learning techniques: Random Forest and LSTM. Our system achieved impressive accuracy, correctly identifying over 93% of cases in one study and 97% in another. This system combines the strengths of both methods, making it effective at both pinpointing key factors and understanding how the disease progresses over time. This promising approach could help doctors diagnose thyroid disease earlier and improve patient care.

INTRODUCTION :

Thyroid disease represents a prevalent health concern, affecting millions of individuals globally. The thyroid, a small butterfly-shaped gland located in the neck, plays a crucial role in regulating various bodily functions through the secretion of hormones. Disturbances in thyroid function can lead to a spectrum of disorders, ranging from hyperthyroidism, characterized by excessive hormone production, to hypothyroidism, where hormone levels are insufficient. The prevalence of thyroid disorders underscores the significance of accurate and timely diagnosis for effective management. In this context, the integration of advanced technologies, such as machine learning and diagnostic imaging, holds promise for enhancing the precision and efficiency of thyroid disease identification and treatment. This paper explores the application of innovative approaches, particularly focusing on leveraging machine learning models, to advance the diagnosis and management of thyroid diseases, aiming to improve patient outcomes and alleviate the burden of this widespread health issue.

PROPOSED METHODOLOGY :

In this project, we tackle the crucial task of thyroid disease prediction through a two machine learning approaches, Random Forest and LSTM algorithms. Our aim is to achieve superior accuracy in identifying individuals who are prone to thyroid disease for tackling it in early stages through clinical approach. As the foundation for our analysis, we used a comprehensive dataset from UCI repository, compiled by Ross Quinlan. This dataset, consisted of data from 3771 individuals and 30 features which are of relevant information to thyroid function. To ensure data integrity, we carefully took care of missing values and eliminated columns, which were not having any significant impact on the individual's health for accurate modeling. First, we used the Random Forest algorithm, known for its ability to handle complex interactions between features. This model served and helpful and cared as a baseline, capturing clearly the patterns within the data. Building upon this baseline performance, we further incorporated a Long Short-Term Memory (LSTM) network. By capturing dependencies within the data, LSTMs excel at modeling sequential patterns, a crucial aspect in understanding the dynamic nature of thyroid function. This both models allowed us to capture both static and dynamic features within the data, improving the predictive accuracy of model Through the process feature engineering and scaling techniques, we optimized the data for effective learning. This data preparation, coupled with the help of the Random Forest and LSTM, we significantly improve thyroid disease predictionThis paragraph incorporates the details you provided and highlights the key aspects of your project.

1) Data Extraction :

When initiating our research project on thyroid disease prediction, our primary focus was on acquiring and processing high-quality data. The dataset, consisting of 3771 individuals with 30 features, was sourced from UCI research. The data preprocessing involved handling null values, removing unnecessary columns, and performing feature encoding. To ensure optimal model performance, we employed feature scaling through normalization. The dataset was carefully curated to enhance precision and achieve robust research

outcomes. The combination of Random Forest and LSTM models was chosen for superior accuracy. The research workflow prioritized data quality, drawing inspiration from best practices in the field.

2) Splitting Data :

A dataset including thousands of people medical record of thyroid reports were stored. We constructed training, validation, and testing data from a csv file containing them. We selected 3771 reliable data points from the whole obtained data set, and the dataset was divided into the training and as well as testing datasets are chosen with the following ratios: 80% and 20%, respectively. The dataset that was obtained before the research activity was used for the further steps of exploration.

3) Building the Model :

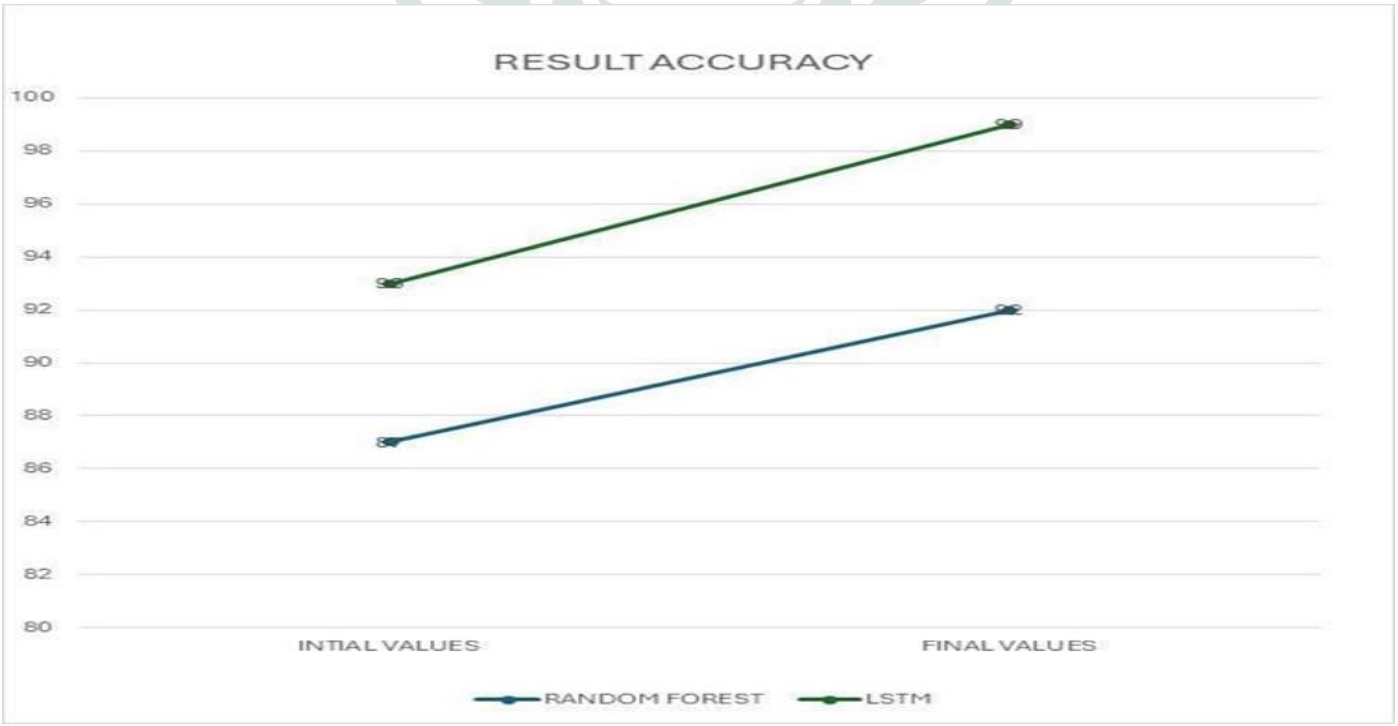
i. Random Forest :

Random Forest acts like a team of various decision trees, each presenting a result on predictions. By compiling their collective results through the averaging, they overcome biases and provide with the accurate predictions, even in tricky situations with missing data. This ability of it makes them a popular choice for many prediction tasks across different fields.Here we could achieve the result of 93% through tuning the parameters through testing and training of data.

ii.LSTM :

The three key components of LSTM networks: 1. Short-Term Memory: Imagine a whiteboard where the network scribbles down relevant information from the current input. This whiteboard, aptly called the cell state, holds what the network needs readily available for the current task. Think of it as the network's temporary workspace, like a student jotting down key points during a lecture. 2. Long-Term Memory: Now picture a filing cabinet in the corner, storing crucial information for the long haul. This is the LSTM's hidden state, a persistent record of everything deemed important throughout the processing sequence. Unlike the whiteboard, the filing cabinet's contents stay put even after the network moves on to new data. This allows the network to tap into past experiences and context, crucial for tasks like predicting text or music composition. 3. Activation Function: Think of this as a gatekeeper, often the sigmoid function, deciding which information flows through the network. It analyzes the current input and the hidden state to assign "scores" between 0 and 1. Scores closer to 1 allow information to pass through freely, while those closer to 0 get blocked. This selective filtering ensures the network focuses on the most relevant data and discards anything that might clutter its path. We achieved a result of 97% by running 100 epochs through the set.

RESULTS :



Comparison of Models :

S.NO	ML MODEL	TRAIN ACCURACY	TEST ACCURACY
1	Random forest	0.87	0.92
2	Long Short term method	0.93	0.97

REFERENCES:

[1] Hosseinzadeh, M.; Ahmed, O.H.; Ghafour, M.Y.; Safara, F.; Hama, H.; Ali, S.; Vo, B.; Chiang, H.S. A multiple multilayer perceptron neural network with an adaptive learning algorithm for thyroid disease diagnosis in the internet of medical things.

[2] Alyas, T.; Hamid, M.; Alissa, K.; Faiz, T.; Tabassum, N.; Ahmad, A. Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach

[3] Das, R.; Saraswat, S.; Chandel, D.; Karan, S.; Kirar, J.S. An AI Driven Approach for Multiclass Hypothyroidism Classification. In Proceedings of the International Conference on Advanced Network Technologies and Intelligent Computing

[4] Razia, S.; SwathiPrathyusha, P.; Krishna, N.V.; Sumana, N.S. A Comparative study of machine learning algorithms on thyroid disease prediction.

[5] Jha, R.; Bhattacharjee, V.; Mustafi, A. Increasing the Prediction Accuracy for Thyroid Disease: A Step towards Better Health for Society