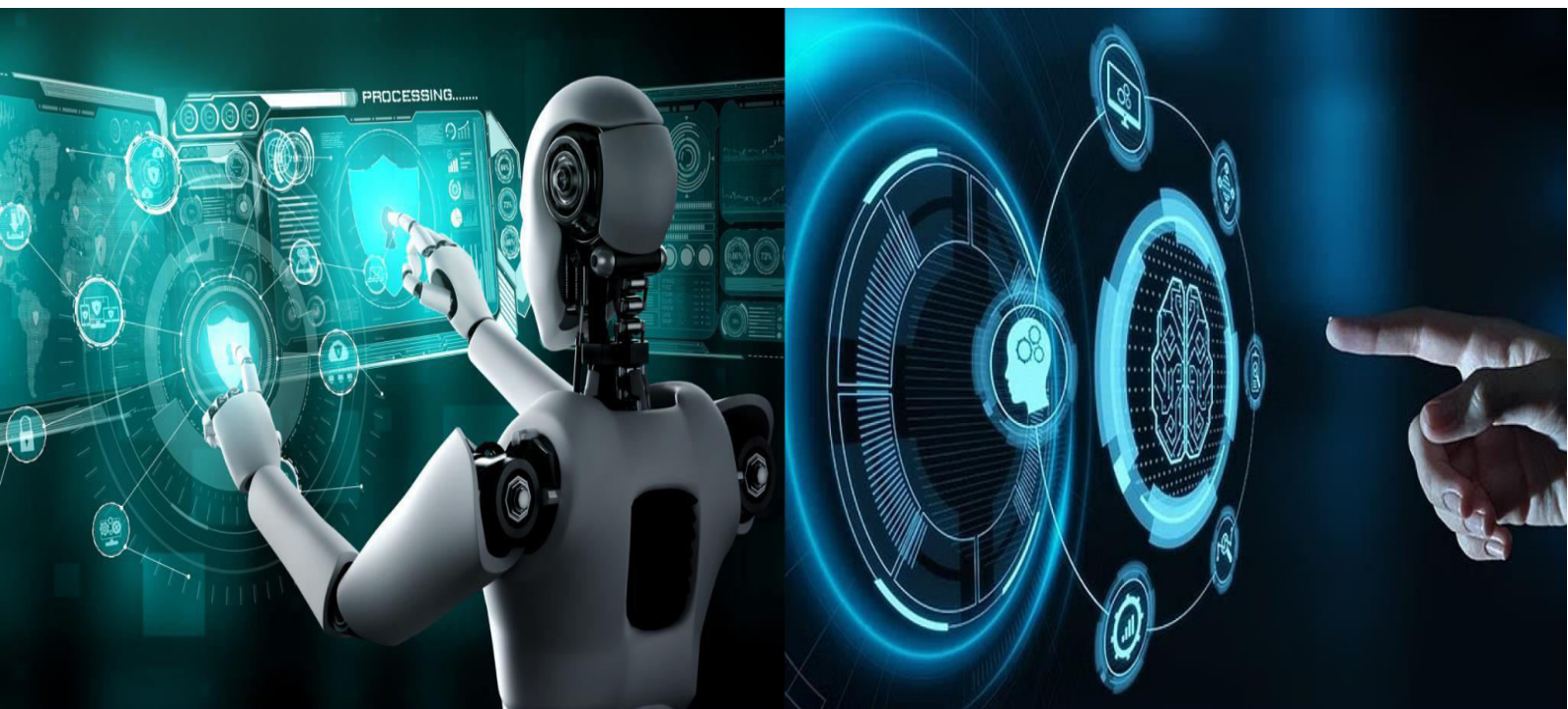


International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





A Machine Learning-Based Framework for EHR-Driven Smart Health Applications

Chandana G O, Santhosh S.G

PG Student, Dept. of MCA, JNN College of Engineering, Shivamogga, Karnataka, India

Associate Professor, JNN College of Engineering, Shivamogga, Karnataka, India

ABSTRACT: Thyroid and diabetes counted amidst the chronic illnesses on the increase globally posing a multitude of challenges to the public health. The discernment of disease at an early stage and its active management are essential to the outcomes of pertaining to the ill person and the healthcare burden. This project brings in online healthcare forecasting system that aims at providing a means of prediction of early risk of diabetes and thyroid ailments by making it available to the masses in a convenient place. Our system uses the Random Forest Classifier, an algorithmic framework model that has been trained on relevant health data sets. The application was created based on the Flask framework, and it allows the patients to input their health parameters and instantly get a prediction concerning their possible health condition. The platform provides tailored healthcare advice and enables one to easily book appointments with health professionals as well as a dashboard through which doctors can facilitate appointments and check the history of predicting patients and patient feedback. This besides prediction. Such a combined solution also works to place greater responsibility on individuals and help health providers make more informed decisions and deliver timely care faster, thus helping shift their favor a preventative approach instead of treating conditions

KEYWORDS: Diabetes, Thyroid, Machine Learning, Random Forest, Predictive Healthcare, Web Application, Flask, MongoDB, Early Detection, Health Management.

I. INTRODUCTION

In today's world to be healthy and don't wait till things turn worse is of utmost importance. Long term illnesses like diabetes and thyroid condition has affected a lot of people all around the world. If these conditions escape diagnosis and handled, within a promptly, they may lead to serious health complications and enormous costs for healthcare systems. This project hopes to use intelligent computer algorithms or machine learning to aid in the early detection of health problems. Consider a world in the term predictive health describes the ability to determine one's risk for illness early-before it escalates or manifests severely. Diabetes is a chronic illness impacting thousands, which, if not properly managed, may results in cardiovascular complications, kidney problems and vision disorders. The way your body burns energy can be impacted if you have a thyroid issue, and it can also lead to many different unpleasant symptoms and health concerns. There is a big call for early diagnosis techniques because these conditions are so common and they can be so deadly serious. Great developments are taking place in the medical field with computers and technology. There are smart devices that monitor your steps, intricate programs allowing the analysis of medical scans, and so on, technology can assist doctors and patients significantly. One form of artificial intelligence, machine learning, is especially practical, in that it can be taught to decipher patterns and make forecasts based on a significant amount of data. The present project brings this approach to the development of a simple albeit powerful tool that any individual can implement on his/her own computer or phone. The present project is one that provides an opportunity to people to test their threat of diabetes and thyroid problems through a simple web page. It cannot replace a doctor; it is simply meant to alert people so that they may seek the attention of the doctor earlier. The system that we have devised:

- Enables users to input the details regarding their health. Uses advanced computer modeling to ascertain their potential be in danger. Gives actionable advice on the cornerstone of the prediction. Allows one to make doctor appointments. Gives physicians a special place to see the information of a patient and leave feedback. In this way, we expect to simplify the task of people recognizing their risks and receiving appropriate care at the appropriate time.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. RELATED WORK

The facilitated interaction between the patient and the physician that is typically absent in the existing healthcare structures focuses more on the treatment of the disease than on identifying risks at an early stage. Existing studies implemented machine learning in single-disease prediction, whereas multi-disease tools are restrained. Ashraf I, Rodriguez CL, Vidal Mazon JL, Chaganti R, De La Torre Dez I, and Rustam F are the authors. 2022,[1] Cancers (Basel) conducts feature engineering in thyroid classification, with unequal classes, and feature selection by FFS, BFE, BiDFE, and extra trees on UCI data. Uddin KMM and others 2024, [2] Science Direct (in press) Compares various base classifiers to a hardvoting ensemble to classify the presence of a thyroid disorder; compares five ML models along with hardnaming ensemble to identify the best result; is concerned with both model comparison and selection and model selection in practice; abstract shows that the ensemble produces better results, but there is still little information on the components of the dataset used and the absence of external validation. Mak I 2024, [3] BMC Med Inform Decis, The propose explainable pipeline is based on a tuned LightGBM classifier and SMOTE-NC imbalancing. Kim K, Kim KS, Kim HK, 2022 [4] Frontiers in Bioscience-Landmark Review of ML for thyroid diseases across different labs and radiology suggests better data quality, multiclass framing, explainability and external validation. It also brings about the issue of small heterogeneous data, a lack of standardization, and generalizability issues. S. Akter and H. A. Mustafa In 2024,[5] the paper in PLoS one compares and interprets the models to classify thyroid using a balance approach and explainable techniques. It also contrasts ML methods in terms of cluster-based resampling, SHAP-style interpretation and cross-validation and concludes that it compares favorably with boosted tree and other classifier. It also states that its limitations of generalization of modified forms of binary/multiclass UCI variants to the larger clinical populations are mentioned. Sharma T, et al. 2021, [6] Advances in Visual Computing for Industry, Biomedicine, and Artistic endeavors Review of diabetes diagnosis using ML/DL, the publicly available datasets, features and model types used, and the patterns of their performance; the barriers to the deployment, including a lack of interpretability and data heterogeneity, and the need to standardize the evaluation and report it transparently to enable the clinical uptake. Hamadi SS, Mahmood SA. [7] In this study, Informatica will compare the base models (SVM, KNN, GNB, LR, DT, RF, NB) and ensembles (voting, bagging, AdaBoost, gradient boosting, and stacking) to predict thyroid disorders of diabetic patients in 2025. The company emphasizes clear that its ensemble strategies are performing well, puts the findings in context of the preceding thyroid ML studies, and points out that validation of the findings in different diabetic subcohorts is necessary. 2022, Sharma K, Wagai GA, and Firdous S. [8] J Family Med Prim Care Survey of ML approaches to diabetes risk prediction across common datasets (e.g., Pima), algorithms (k-NN, SVM, RF, LR), and preprocessing; highlights the prominence of larger and more inclusive clinical cohorts, coupled with transparent reporting, to generalize risk prediction models; and it notes variability in performance based on feature engineering, data imbalance. L. Fregoso-Aparicio and colleagues 2021,[9] Diabetology & Metabolic Syndrome Systematic review of ML/DL to predict type 2 diabetes; compares 18 model families, with tree-based methods often performing best; identifies problems: inconsistent feature sets, heterogeneity in datasets and reporting gaps, which impair selection of models; recommend standardized benchmarks and interpretability to overcome obstacles to clinical implementation. Healthcare (Basel) (PMCID) integrates healthcare characteristics, standard preprocessing, comparisons of various machine learning and neural network settings to forecast diabetes onset, improved predictive data with specific structures, warns against overfitting, and demands increased studies opportunities, such as larger and multicentric datasets, and explainability to transform clinical decision support.

III. PROPOSED ALGORITHM

Our healthcare predictive method is developed with the various parts in tandem with each other like a well oiled machine. There is a web component in which users interact, a smart machine learning component that does the predictions, and then there is the storage component which stores all the data.

3.1 Data Collection and Preprocessing

Our intelligent models must be taught with the current information that they already possess, before they can be able to make a prediction. We mainly employed the two groups data, one on thyroid ailments and the other on diabetes. 1. Reading the Data: The data was read using the Pandas program as it converts the CSV files into data. 2. Data Cleaning and Preparation: The computer is more inclined towards handling numbers than words. Consequently, the text answers like Yes or No, Male or Female had to be changed into numbers (e.g. 1 or 0). Another thing we did with the thyroid data was optimize some additional categories such as Weight Change, Fatigue, etc, into numbers using a LabelEncoder. This step is essential to affirm that our models can grasp the information. 3. Scaling Features: These are the numbers



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Age (say 50), and GlucoseLevel (say 150). The maximums of these figures are very diverse. To make all the information equal to the models, we apply StandardScaler to change the nature of all the numbers. This is owing the model will learn better when the domain of numbers is similar. 4.Splitting Data: This involves dividing the given data into two subsets i.e. a training set, on which the model learns, and the evaluation set, wherein we check that a model learned effectively. We devoted 20 percent on testing and 80 percent on training.

3.2 System Architecture

The entire system is designed in layers: • Web Application Layer (Flask): The users make use of this layer. It was built using a popular site building tool: Flask. The first connection (called the front-end) entails the displaying of pages, the ability of the user to either log in or register and the reception of the health information they write. The Machine Learning Layer This little algorithm is the brain of our system. Using special computer programs (models), the health data is analysed and predictions suggested. Thyroid and diabetes predictions are two features that we implement the Random Forest type of models. This is where all the metrics is secured- MongoDB Database layer. We also employ MongoDB which is beneficial in storing any type of information such as user accounts, results on predicted information, and appointment data.

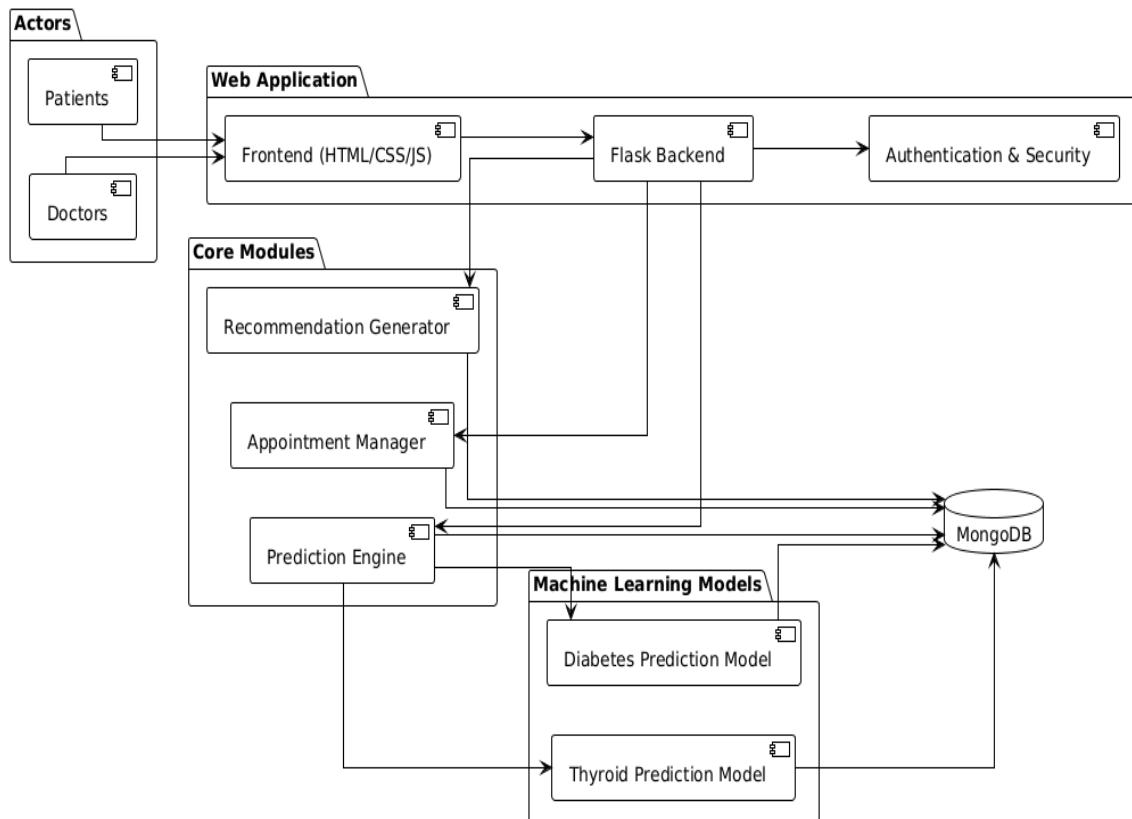


Fig 3.2 system architecture

3.3 Machine learning models

We applied Classification Random Forest Algorithm in diabetes and thyroid disease prediction. Random Forest is the ensemble classification approach that combines multiple trees to enhance the predictive accuracy and less overfitting. The algorithm orchestrates a set of decision trees, each systematically constructed from a different subset of training data and features, and the many trees are combined so that their predictions are decided by a majority voting rule. The Random Forest algorithm can be mathematically represented as:

$$y^{\wedge} = B1b = 1\sum BTb(x)$$



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Where y^{\wedge} is the final prediction, B is the no of trees, and $T_b(x)$ is the prediction from the b -th tree. For feature selection and importance calculation, Random Forest uses the Gini impurity measure:

$$Gini = 1 - i = 1 \sum n p_i^2$$

Where p_i is the probability of class i in the node.

Data Preprocessing and Feature Engineering

Several crucial steps are integral to the be run through machine learning. To apply on categorical variables, we use the Label Encoding to transform text to numerical value. LabelEncoder also gives a mapping between the categorical of a value and the integer in the dataset, so this will be consistent.

For numerical features, we apply StandardScaler normalization to guarantee that all features equally influence the model. The standardization formula is:

$$z = \frac{x - \mu}{\sigma}$$

Where z is the standardized value, x is the original value, μ is the mean, and σ is the standard deviation.

3.4 Prediction Pipeline

This is the process when a new patient submits their information on their store: 1. The system takes all the entered symptoms and health numbers, and 2.Encode and Scale: The system encodes symptoms that have words, i.e., when the data respond to variables, the Label Encoders that had been trained are applied to convert them into numbers. Then all the figures are scaled by the saved Standard Caler. This ensures the new data resembles the data the model has been trained on.3. Display Result: The prediction that the model made like, "Diabetic," or, "Normal Thyroid," is shown to the user. Further, we show what we call a confidence level that describes how certain the model is with its prediction.

3.5 Web Interface (Flask) and Database (MongoDB)

All user interactions with Flask web application are managed User Management: Users are allowed to create log-ins and new accounts. Prediction Forms: Purposeful forms on entry of diabetes and thyroid-related health state. The confidence and prediction results are well-represented. Recommendations: Common-condition-based, individualized suggestions on things such as exercise and diet. Appointment Booking: One has the ability to formulate appointments with doctors using this system. Doctor Dashboard: A safe place where doctors can view all of their appointments, patient prediction histories and provide an update and/or instructions. All of this information: user profiles, prediction results, set appointments, and even feedbacks of doctors are stored in the MongoDB, so it is easy to manage and access all this data.

IV. RESULT AND DISCUSSION

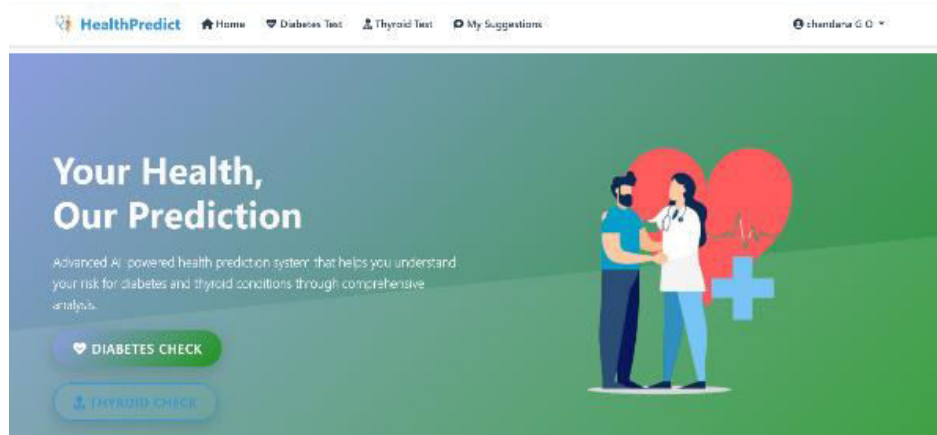


FIG 4.1 HOME PAGE



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

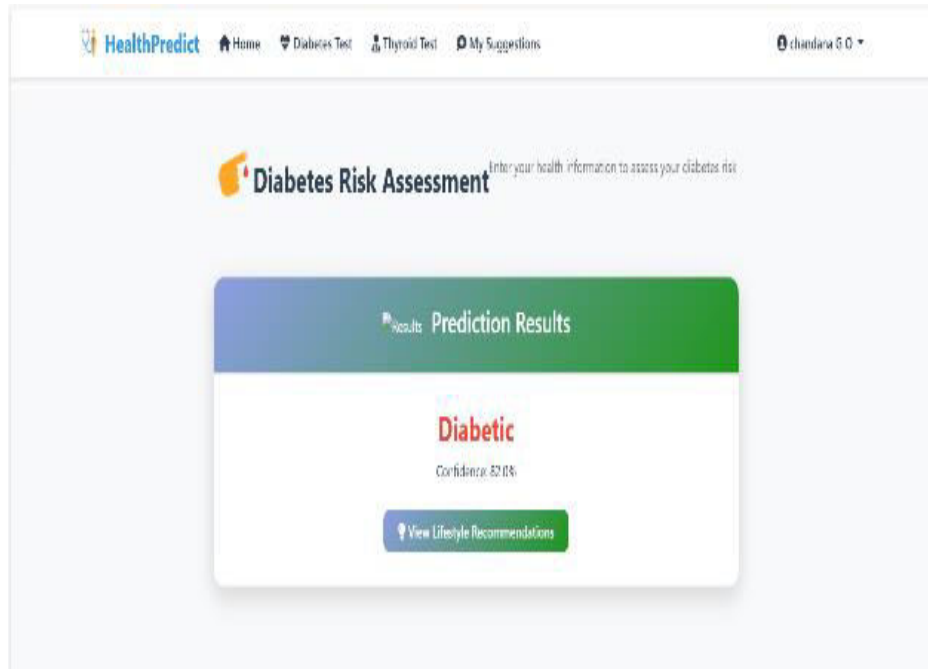


FIG 4.2. PREDICT THE DIABETES

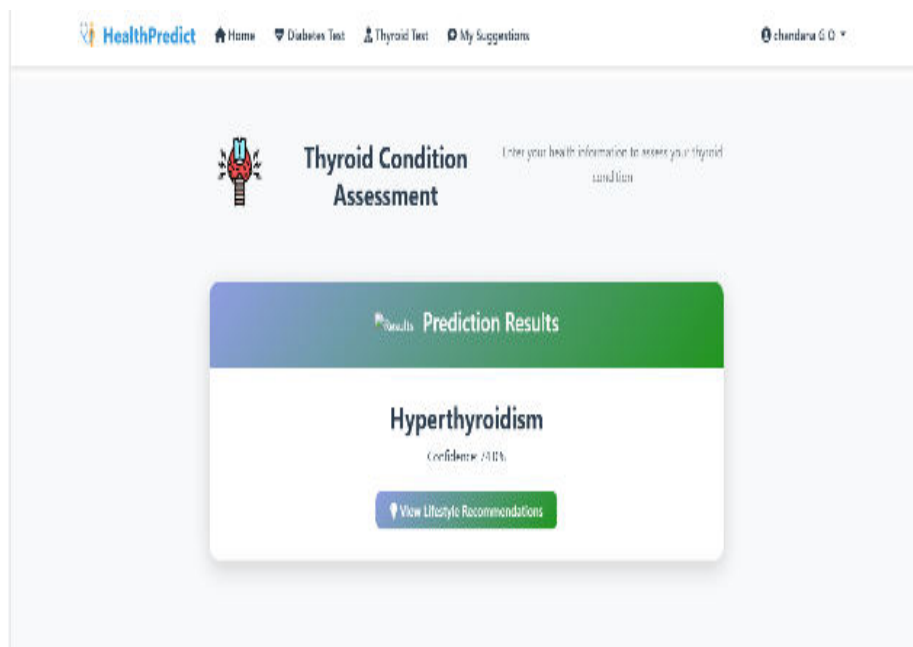


FIG 4.3. PREDICT THE THYROID

The project was successfully able to develop a working web app, which utilizes machine learning to assist in predicting diabetes and thyroid conditions. The Random Forest models that we used performed extremely well based on the test. As an example, the diabetes model had Accuracy Score that was frequently above 90% i.e., it predicted the presence or absence of diabetes quite often. The thyroid model also performed accurately in the differentiation of hypothyroidism, hyperthyroidism and normal.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

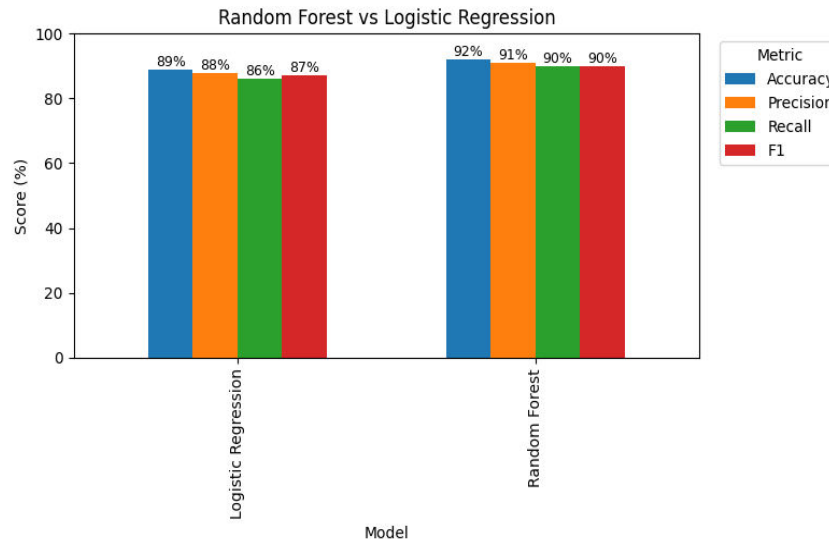


FIG 4.4. COMPARISON GRAPH

The comparison between Logistic Regression and Random Forest indicated that Random Forest outshone the other classification method. Logistic Regression was an easy model that parted the data along a straight course and had an accuracy of 89%. Random Forest utilizing multiple decision trees to create more accurate predictions had 92 percent of accuracy with better precision, recall, and F1-score. This means that Logistic Regression is the more simple and easy to interpret compared to Random Forest which is more effective in recognizing a pattern in the data.

The more detailed classification_report produced by our model training told us more. In the case of the primary health classes, its metrics were distinguished by elevated values of precision and recall. Recall will inform us of how many of the real positive cases were found in our model, whereas precision will tell us how many of the positive predictions were accurate. The large values in these regions certify the validity of the models. The confusion matrix also clearly indicated that very few errors were made especially in the comparison of healthy individuals to sick individuals or sick individuals to those that were healthy. To the user perspective, the web application is user-friendly. The health data entry forms are simple to comprehend and the prediction outcomes and strength is shown instantly. This enables users to estimate their risk fast enough, and in an easy way without necessarily possessing a lot of medical understanding.

V. FUTURE ENHANCEMENT

Though this project has provided an excellent basis of a practical healthcare prediction system, a plethora of potentials in the future to strengthen it in capability. As a way of making the system more comprehensive, it may be extended to predict other notable chronic disorders including heart disease, kidney disease and even cancerous disease of some types. The data would be reconciled in real time between sources such as smartwatches, and electronic health records (EHR), which would lead to more personal and precise health monitoring. Technologically, one can explore new models, such as deep learning which can be used to identify some hidden trends in bigger data sets, and make more accurate predictions. The platform can be enhanced with personalized dashboards with the trends across time, medications notifications, and progress reports to users. Furthermore, the software would be turned into a more holistic personal health management program with such features as interactive diet and exercise plans that would meet the requirements of a given user.

VI. CONCLUSION

In summary, a web-based healthcare prediction system that is intuitive, simple and easy to use concerning diabetes and thyroid disorders has been created and implemented successfully in this project. Through the capability of machine learning, namely the Random Forest Classifier, we have developed a tool that allows people to have an early warning of their health hazards. The potential to predict high and low precision results on a range of health data and



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

demonstrate reliance on the findings through the predictive modeling reveals a high level of impact to the system of proactive health care management. This provides the users with the opportunity to make the effort to lead a healthier life. Moreover, the incorporated appointment booking system helps in ensuring prompt medical consultation since the engagement of patients with medical practitioners will be efficient. This project is an important contribution to the goal of adopting predictive healthcare by ordinary users. It shows how technology may be utilized successfully to help in the early diagnosis of the disease, motivate prevention, and enhance consumer- physician communication.

REFERENCES

1. Rustam F, De La Torre Díez I, Chaganti R, Vidal Mazón JL, Ashraf I, Rodríguez CL. Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques. *Cancers (Basel)*. 2022;14(16):3914. doi:10.3390/cancers14163914
2. Uddin KMM; et al. An ensemble machine learning-based approach to predict thyroid disorders. 2024; journal article on ScienceDirect, article in press with methods summarized in abstract
3. Raza A; Xu X; et al. Enhanced interpretable thyroid disease diagnosis by explainable AI. *BMC Med Inform Decis Mak*. 2024; article number listed by the journal
4. Lee KS, Kim HK, Kim K, et al. Machine learning on thyroid disease: a review. *Frontiers in Bioscience-Landmark*. 2022;27(3):101. PDF available with full pagination in the journal's PDF
5. Akter S, Mustafa HA. Analysis and interpretability of machine learning models to classify thyroid disease. *PLoS One*. 2024;19(5):e0300670. doi:10.1371/journal.pone.0300670
6. Sharma T, et al. A comprehensive review of machine learning techniques on diabetes detection. *Visual Computing for Industry, Biomedicine, and Art*. 2021;4:30. Article number 30
7. Mahmood SA, Hamadi SS. Ensemble Machine Learning Algorithms for Predicting Thyroid Disorders in Diabetic Patients: A Comparative Analysis. *Informatica*. 2025;49:173–184. PDF indicates pages 173–184 in vol.49
8. Classification of thyroid diseases using machine learning. *International Journal of Health Sciences (ScienceScholar)*. 2022; PDF article with authors and pagination indicated on PDF
9. Firdous S, Wagai GA, Sharma K. A survey on diabetes risk prediction using machine learning approaches. *J Family Med Prim Care*. 2022;11(11): article with PMCID provides details; includes multiple method comparisons
10. Machine Learning Approach with Harmonized Multinational Datasets for enhanced prediction of hypothyroidism in patients with type 2 diabetes. 2024; PMCID article with AUROC/NPV; clinical decision support framing.
11. Thyroid Disease Classification using Machine Learning (conference). *E3S Web of Conferences*. 2023; ICMED-ICMPC 2023 proceedings PDF with authors and pagination Fregoso-Aparici.
12. o L, et al. Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. *Diabetology & Metabolic Syndrome*. 2021;13:148. Article number 148
13. Enhancing thyroid disease prediction and comorbidity management (2025). *ScienceDirect/Springer 2025 article summarizing selective feature extraction benefits and comorbidity aspects*.
14. Semi-Supervised Machine Learning Approaches for Thyroid Disorder Diagnosis. *International Journal of Interactive Multimedia and Artificial Intelligence*. 2024;8(7): paper PDF with authors and article details.
15. Chou C-Y, et al. Predicting the Onset of Diabetes with Machine Learning. *Healthcare (Basel)* or related open-access venue via PMCID. 2023; includes multipleneuralnetworks comparisons.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



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