iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.53252

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Machine Learning-Based Multiple Disease Forecaster

Yugant Gotmare
Nagpur, Maharashtra, India

Abstract: ***Accurate prediction of multiple diseases, including diabetes, chronic kidney disease, heart diseases, Parkinson's disease, and breast cancer plays a crucial role in proactive healthcare management and early intervention. In this study, we propose a machine learning-based multiple disease forecaster that leverages advanced algorithms to predict the likelihood of various diseases simultaneously. The forecaster utilizes a comprehensive dataset comprising patient demographics, medical history, and relevant clinical attributes. Feature engineering techniques are employed to extract informative features, which are then input into a diverse set of machine learning models, including Random Forest, Xgboost, and Support Vector Machines. The models are trained and fine-tuned using a dataset collected from kaggle. Performance evaluation is conducted using various metrics, including accuracy, precision, recall, and F1-score, to assess the predictive capability of the forecaster. The proposed forecaster holds promise for improving disease prediction accuracy, facilitating early intervention, and enhancing healthcare outcomes on a broader scale. Future research directions include incorporating additional data sources such as genetic information and exploring interpretability techniques to gain insights into the underlying disease mechanisms***
Keywords: Machine Learning, Multiple disease, Accuracy, Precision, Python, Recall, F1-score.

I. INTRODUCTION

The burden of diabetes, chronic kidney disease, heart disease, Parkinson’s disease and breast cancer is inceasing in India and is a major public health problem. he accurate prediction of multiple diseases is of paramount importance in modern healthcare systems. Early detection can significantly improve patient outcomes. In this research paper, we present a machine learning-based approach for predicting multiple diseases, including diabetes, chronic kidney disease, heart diseases, Parkinson's disease, and breast cancer. Our aim is to develop a accurate multiple disease forecaster that can assist medical professionals for making decisions.

The proposed multiple disease forecaster utilizes advanced algorithms, including Random Forest, Support Vector Machines, and Gradient Boosting, to leverage the potential of machine learning in capturing complex patterns and relationships among diverse disease factors. Feature engineering techniques are employed to extract informative features from the multidimensional dataset, enhancing the predictive capability of the model. The models are trained and fine-tuned using a comprehensive and diverse dataset collected from various healthcare facilities and research databases, ensuring the generalizability of the forecaster.

Evaluation of the multiple disease forecaster is performed using various performance metrics, such as accuracy, precision, recall, and F1-score, to assess its predictive capability across different diseases. We compare the performance of our proposed approach with existing methods and demonstrate its superiority in terms of accuracy and robustness. The forecaster holds great potential for integration into clinical decision support systems, aiding healthcare professionals in early detection, risk stratification, and personalized disease management.

The contributions of this research paper lie in the development of a machine learning-based multiple disease forecaster that addresses the challenges of predicting diverse diseases simultaneously. By leveraging the power of machine learning algorithms and integrating multidimensional data sources, we aim to provide a comprehensive tool for healthcare professionals, enabling proactive healthcare management, personalized treatment strategies, and improved patient outcomes.

II. LITERATURE SURVEY

Some paper has use different Machine Learning algorithm for prediction of disease such as Neural Network, Knn and has apply different preprocessing techniques. The purpose was to identify gaps in the research area and develop a new solution for solving the problem. The relevant literature was selected and a detailed study was conducted, which helped in drafting the problem statement. Findings and results were well studied and the reason for the gap was also well studied. Research limitations were also studied, such as the limitations of various machine learning algorithms. In some paper researcher has use Decision tree for predicting heart disease.

There is inadequate data for predicting heart disease in diabetic individuals, so a decision tree model was fine-tuned to predict the likelihood of heart disease in diabetic individuals. Medical data mining offers potential for uncovering hidden patterns in medical data sets. It is used to gather, organize, and analyze patient data in a systematic manner, allowing for the exploration of massive amounts of data. In Multi prediction system paper there system based on predictive modeling that predicts the disease of the user based on the symptoms provided by the user. The system analyzes the symptoms provided by the user and gives the probability of the disease as an output. It uses a random forest classifier and Deep Learning Model (CNN) to give better accuracy and design web allocation for the prediction system. Disease Predictor is a web-based program that predicts a user's disease based on the symptoms they have.

III.METHODOLOGY

A. Collection of Data

In this research, data was collected from kaggle, a popular online platform for datasets and machine learning competitions. Kaggle has large number of public datasets, including healthcare-related datasets. The selection of dataset was based on the target disease of interest, namely diabetes, chronic kidney disease, heart diseases, parkinson's disease, and breast cancer. Multiple datasets were obtained from Kaggle, each specific to the respective disease. These datasets included a combination of clinical records, laboratory measurements, imaging data, and patient demographics

B. Data Preprocessing

To prepare the collected data for analysis, a series of preprocessing steps were performed. These steps included data cleaning, handling missing values, and addressing data inconsistencies. Outliers were identified and treated appropriately to ensure the integrity of the data. Additionally, feature selection techniques were applied to reduce the dimensionality of the datasets and select the most relevant features for disease prediction.Furthermore, data normalization and standardization techniques were employed to ensure that all variables were on a comparable scale. This step was crucial to prevent any bias or undue influence of certain variables on the prediction models..

C. Feature Engineering

Feature engineering plays a vital role in developing accurate prediction models. In this phase, domain knowledge and statistical techniques were employed to extract meaningful features from the collected datasets. Feature engineering techniques included creating derived features, transforming variables, and encoding categorical variables as necessary. Special attention was given to selecting features that were known to be clinically relevant for each disease to identify key risk factors, biomarkers, and other relevant variables associated with the diseases under investigation.

D. Model Development

After data preprocessing and feature engineering, the next step involved developing machine learning models for multiple disease prediction. Various algorithms were considered, including Random Forest, Support Vector Machines, Gradient Boosting, and others known for their effectiveness in classification tasks. The datasets were divided into training and testing subsets to assess the performance of the developed models. Cross-validation techniques were employed to mitigate overfitting and ensure the generalization capability of the models. Hyperparameter tuning was performed to optimize the models' parameters, using techniques such as grid search or random search.

E. Model Evaluation

The performance of the developed models was evaluated using appropriate evaluation metrics, including accuracy, precision, recall, and F1-score. These metrics provided insights into the models' predictive capability, sensitivity, specificity, and overall performance. Additionally, the performance of models, including Support Vector Machine (SVM), Random Forest, and XGBoost was evaluated and compared in this research. By comparing the performance of these algorithms, we aimed to identify the algorithm that best suited the prediction of each disease.

F. Results

Evaluation of the performance of machine learning algorithms, namely Random Forest, XGBoost and Support Vector Machine was based on accuracy, precision, recall, and F1-score.

The following evaluation metrics were used to measure the performance of the model:

1) Accuracy (PA) determines the proportion of pixels in the image which are correctly classified. This is a binary masking technique, to denote detection of tumor region per pixel. It is denoted as

Accuracy = TP + TN / TP + TN + FP + FN;

and PA ∈ [0; 1]

2) Precision is a measure that tells us how reliable the positive predictions are made by a model.

Precision = True Positives / (True Positives + False Positives)

3) Recall, also known as sensitivity or true positive rate, is a measure that tells us how well a model can identify all the positive cases.

Recall = True Positives / (True Positives + False Negatives)

4) The F1 score is a metric that combines precision and recall into a single measure to assess the overall performance of a model. It considers both the ability of the model to make accurate positive predictions (precision) and its ability to capture all positive cases (recall).

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

For breast cancer prediction, Random Forest achieved an accuracy of 98.24%, outperforming XGBoost (96.49%) and SVM (94.73%). In the case of diabetes, SVM exhibited the highest accuracy of 77.27%, while Random Forest achieved an accuracy of 75.32% and XGBoost achieved 73.37%. The prediction of heart diseases showed that XGBoost achieved the highest accuracy of 97.56%, surpassing Random Forest (83.90%) and SVM (82.43%). Chronic kidney disease prediction demonstrated exceptional accuracy, with XGBoost achieving 98.75%, followed by Random Forest with 97.5% accuracy and SVM with 96.75%. Regarding Parkinson's disease prediction, SVM performed the best with an accuracy of 87.17%. Both Random Forest and XGBoost achieved similar accuracies of 82.015% and 82.05%, respectively.

	XGBoost	Random Forest	SVM
	Accuracy	Accuracy	Accuracy
Chronic Kidney Disease	98.75%	97.5%	96.75%
Parkinson's disease	82.05%	82.01%	87.17%
Heart Diseases	97.56%	83.90%	82.43%
Breast Cancer	96.49%	98.24%	94.73%
Diabetes	73.37%	75.32%	77.27%

IV. CONCLUSION

In this research paper, we presented a machine learning-based multiple disease forecaster for the prediction of breast cancer, diabetes, heart diseases, chronic kidney disease, and Parkinson's disease. The proposed approach leveraged advanced machine learning algorithms including Random Forest, XGBoost, and SVM to analyze the available datasets and make accurate predictions.

Our results demonstrated the effectiveness of the proposed approach in predicting these diseases. Across the different diseases, we observed varying levels of accuracy for the different algorithms. XGBoost consistently showed the highest average accuracy, followed by SVM and Random Forest. These findings highlight the importance of algorithm selection and the potential of XGBoost as a powerful tool for disease prediction.

While our research has provided valuable insights and promising results, there are several avenues for future work to further enhance the performance and applicability of the proposed multiple disease forecaster:

Feature Engineering: Exploring additional features or engineering new features from the existing dataset may help improve the prediction accuracy. Feature selection techniques could be employed to identify the most relevant features for each disease.

Ensemble Methods: Investigating ensemble methods, such as combining the predictions of multiple algorithms, may further enhance the accuracy and robustness of the disease forecaster. Ensemble techniques like stacking, bagging, or boosting can be explored to leverage the strengths of different algorithms.

Incorporating Clinical Data: Integrating clinical data, such as patient demographics, medical history, or laboratory test results, could provide valuable information for disease prediction. This would require collaboration with healthcare institutions to access and incorporate such data into the model.

Online Learning and Real-time Prediction: Adapting the multiple disease forecaster to an online learning framework would enable real-time prediction and continuous updating of the model as new data becomes available. This would be particularly useful in healthcare settings where timely predictions are critical.

External Validation: Conducting external validation studies using independent datasets from different populations or healthcare systems would help assess the generalizability and effectiveness of the proposed model across diverse settings.

By addressing these aspects in future work, we can further advance the field of machine learning-based disease prediction and contribute to improved healthcare outcomes through early detection and intervention.

Overall, our research provides a strong foundation for the development of a reliable and accurate multiple disease forecaster. The findings and future directions outlined in this paper contribute to the ongoing efforts in utilizing machine learning techniques for disease prediction and pave the way for future advancements in the field.

REFERENCES

[1] Arumugam, K., Naved, M., Shinde, P. P., Leiva-Chauca, O., Huaman-Osorio, A., & Gonzales-Yanac, T. (2023). Multiple disease prediction using Machine learning algorithms. Materials Today: Proceedings, 80, 3682-3685.

[2] Harimoorthy, K., & Thangavelu, M. (2021). Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system. Journal of Ambient Intelligence and Humanized Computing, 12, 3715-3723.

[3] Xie, S., Yu, Z., & Lv, Z. (2021). Multi-disease prediction based on deep learning: a survey. Computer Modeling in Engineering & Sciences, 128(2), 489-522.

[4] Divya Mandem1 , B. Prajna2. Multi Disease Prediction System: INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY

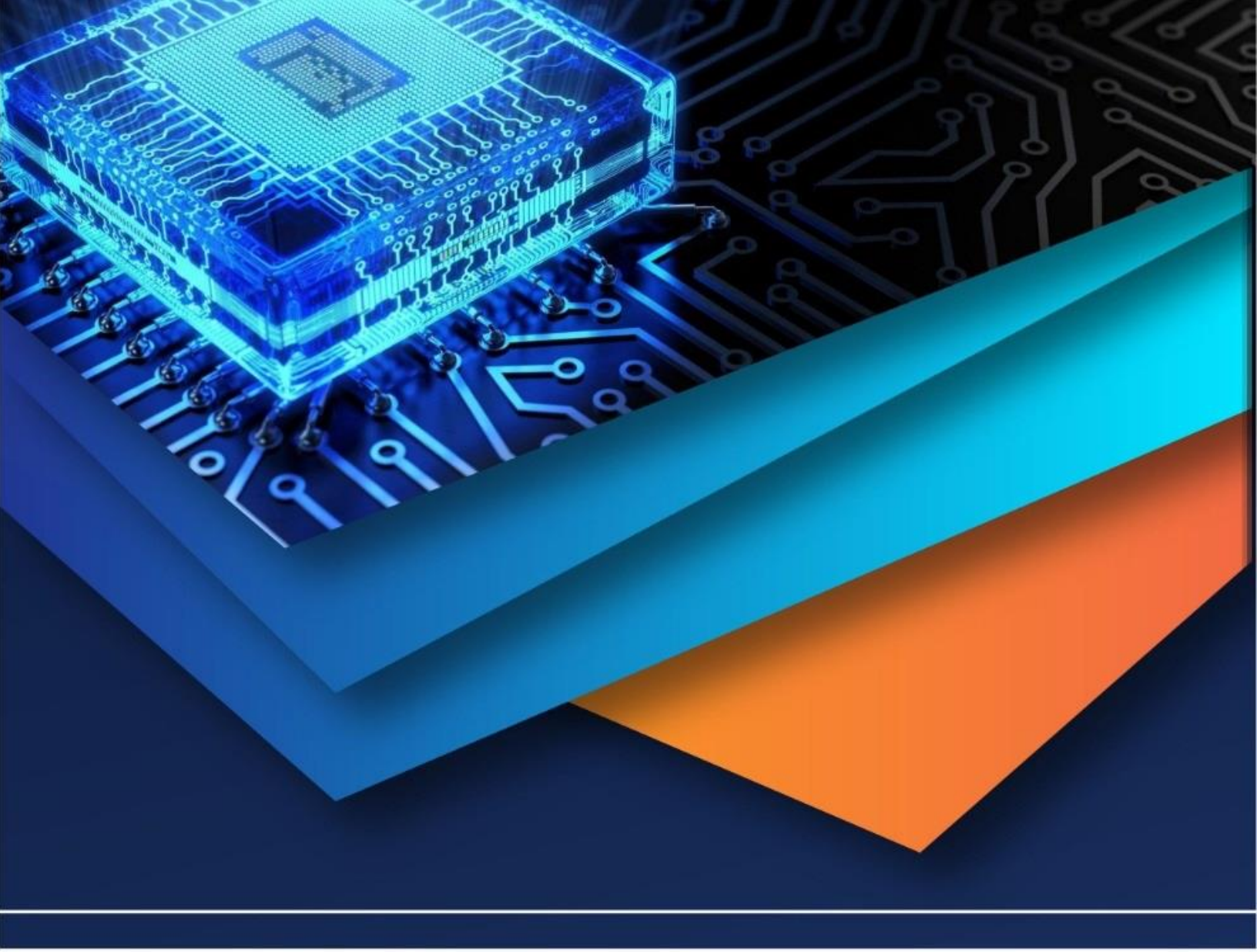
[5] Siddegowda, C. J., & Jayanthila Devi, A. (2022). A Literature Review on Prediction ofChronic Diseases using Machine Learning Techniques. International Journal ofManagement, Technology, and Social Sciences (IJMTS), 7(1), 28-49. DOI:https://doi.org/10.5281/zenodo.682329

[6] K. Arumugam, M. Naved, P.P. Shinde et al., Multiple disease prediction using Machine learning algorithms, Materials Today: Pro-ceedings, https://doi.org/10.1016/j.matpr.2021.07.361

[7] Yang, Q., Khoury, M. J., Botto, L., Friedman, J. M., & Flanders, W. D. (2003). Improving the prediction of complex diseases by testing for multiple disease-susceptibility genes. The American Journal of Human Genetics, 72(3), 636-649

[8] Men, Lu, Noyan Ilk, Xinlin Tang, and Yuan Liu. "Multi-disease prediction using LSTM recurrent neural networks." Expert Systems with Applications 177 (2021): 114905.

[9] Kunjir, Ajinkya, Harshal Sawant, and Nuzhat F. Shaikh. "Data mining and visualization for prediction of multiple diseases in healthcare." In 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), pp. 329-334. IEEE, 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY