NIPES

Journal of Science and Technology Research

Journal homepage: www.nipesjournals.org.ng

Disease Prediction Using Random Forest Machine Learning Algorithm

Omankwu, Obinnaya.C.B; Osodoeke, Efe Charlse; and Ubah, Valentine Ifeanyi
Department of Computer Science, Michael Okpara University of Agriculture, Umudike
saintbeloved@yahoo.com

Article Info

***Keywords*:** *Machine Learning Algorithm, Random Forest, Chronic Disease*

Received 09 February 2024
Revised 02 March 2024
Accepted 19 March 2024
Available online 21 April 2024

https://doi.org/10.5281/zenodo.11005308

Abstract

The growing significance of machine learning in healthcare and disease prediction is underscored by its application in the "Disease Prediction" method, which employs predictive modeling to analyze user-input symptoms and forecast potential ailments. Automated disease prediction systems offer a promising solution to the challenges of accessing timely and cost-effective healthcare, particularly for individuals residing far from medical facilities. Leveraging data mining techniques, these systems assess a patient's risk level based on symptoms, facilitating early detection and management of chronic diseases like heart disease and diabetes. Machine learning plays a pivotal role in this domain, empowering predictive models to analyze vast healthcare datasets efficiently. Despite the advancements, there's a need for comprehensive studies integrating machine learning for disease prediction, especially regarding chronic conditions. The proposed framework integrates structured and unstructured data, employing the Random Forest algorithm for accurate predictions. Results demonstrate high accuracy across various diseases, showcasing the potential of machine learning in enhancing healthcare outcomes. However, challenges like dataset biases and overfitting necessitate future research focusing on larger, more representative datasets and advanced modeling techniques. Collaborative efforts between stakeholders can drive the adoption of machine learning-driven predictive models in clinical settings, ushering in a new era of proactive and personalized healthcare.

1. Introduction

If someone is currently ill, they must see a doctor, which is a costly and time-consuming process. It could also be difficult for the patient if they reside distant from medical services because the Disease is unknown [1]. Therefore, if the aforementioned procedure could be carried out using automated software that saves time and money, it might be better for the patient and the process would go more smoothly [1]. Certain Heart Disease Prediction Systems use data mining techniques to determine the patient's risk level. Condition Predictor is an online program that determines a user's condition based on their symptoms [1]. Data sets for the Disease Prediction system have been collected from a variety of health-related websites [2]. Condition Predictor allows the user to determine the likelihood of a condition by using the given symptoms. Learning new things piques people's inherent curiosity, especially in light of the everyday rise in internet usage [3]. When an issue comes up, individuals usually want to look it up online. Hospitals and physicians have less internet connectivity than the general public. People who are ill don't have a lot of options. People might so

gain from this system. A chronic Disease is one that lasts a long time or heals slowly. Many chronic Diseases are not curable; instead, they can only be managed with regular care. Like any other nation, India is going through significant social and economic changes that are causing a notable rise in the incidence of cardiovascular disease [4].

Many affluent countries, including India, are facing a range of chronic ailments, mainly diabetes and heart disease, which might have significant effects on global health, security, and the economy. The rapidly expanding urban and economic landscape of today has given rise to a wide range of lifestyles. Every nation is concerned about the one-third of the population that is afflicted with a chronic Disease nowadays. For the sick, receiving care for chronic conditions is more difficult and expensive [5]. The medical field gathers and processes enormous amounts of datasets related to chronic diseases, and data mining aids in the early detection of Diseases. Diabetes, Parkinson's disease, Alzheimer's disease, liver disease, and cardiovascular disease are among the most costly diseases to be diagnosed [6].

Since only those with the means to do so may benefit from it, providing the finest care possible for all patients is a major issue in the medical and healthcare industries. Despite the massive amount of available healthcare data, it isn't being mined more efficiently or consistently to uncover hidden information needed to make informed decisions. The suggested architecture makes use of data mining techniques to early detect chronic diseases [6].

Machine learning is the process of teaching computers to perform better by using examples or historical data. The study of computer systems that gain knowledge from data and experience is known as machine learning [5]. Training and testing are the two phases of the machine learning algorithm. For a long time, machine learning has made it difficult to forecast diseases based on a patient's medical history and symptoms [7].

Machine learning technology provides a strong foundation for solving healthcare issues in the medical area [7].

The goal of this paper is to create the framework and system that will enable end users to predict chronic Diseases without visiting a physician or doctor for a diagnosis. To use a range of machine learning models and patient observation techniques to diagnose various ailments. There is no established mechanism for handling text and structured data. Both structured and unstructured data will be considered by the proposed system. Machine learning can improve prediction accuracy.

The knowledge gap is the absence of a comprehensive and long-term study that integrates machine learning methods for Disease prediction in healthcare, particularly concerning chronic diseases like diabetes, heart disease, and others. Put another way, even though machine learning is becoming more and more important in healthcare and Disease prediction, there is still a dearth of material in the literature about the development and evaluation of predictive models for chronic diseases using both structured and unstructured healthcare data. Further study could improve Disease prediction models' accuracy and use in real-world clinical settings, as existing studies often lack comprehensive techniques, comprehensive evaluation measures, and trustworthy validation procedures.

1.1. Literature Review

The mining technique is the best medical diagnosis as investigated [8]. The authors of this study compared Nave Bayes to five different classifiers: LR, KStar (K*), Decision Tree (DT), Neural Network (NN), and a basic rule-based approach (ZeroR) [8]. The effectiveness of each method was evaluated using fifteen real-world medical situations from the UCI machine learning collection [9]. Nave Bayes' prediction accuracy results are superior to those of other approaches, as evidenced by NB's superior performance in 8 out of the 15 data sets in the trial. It was discovered that treating

chronic Disease globally is inefficient in terms of both time and financial resources [9]. Thus, the goal of the study was to estimate the authors' potential risk of sickness.

This was accomplished with CARE, which only uses ICD-9-CM numbers and a patient's medical history to determine the probability of a diagnosis. CARE uses a combination of collective filtering approaches and clustering based on the medical history of each patient as well as those of similar persons to predict the highest disease risks for each patient [10]. The iterative version of ICARE, which applies ensemble principles for greater efficiency, has also been specified by the authors.

These cutting-edge technologies don't require significant expertise and can predict a broad spectrum of medical diseases in a single run [11]. Thanks to ICARE's superior potential risk coverage, more precise early warnings for thousands of Diseases, several years in advance, are now achievable. When used to its full potential, the CARE system can investigate a wider range of Disease causes, present original research, and promote discussion on early identification and prevention [12]. A survey of current approaches for information discovery in databases using data mining techniques that are employed in today's medical research, notably in Heart Disease Prediction were reviewed [12]. Several tests have been carried out on the same dataset to evaluate the effectiveness of predictive data mining methods. The findings show that Decision Tree outperforms other predictive techniques, with Decision Tree accuracy occasionally matching that of Bayesian classification. Other less effective predicting methods are KNN, Neural Networks, and Cluster-Based Classification [13].

The Decision Tree Algorithm, which compares user-provided data to a preset set of values, was used to predict cardiac diseases in a study by Adam Pattekari and Asma Parveen. Because of this study, patients were able to provide simple information that was matched to data and heart disease was predicted [14]. A scholar examined the various heart-related conditions using medical data mining approaches such association rule mining, grouping, and clustering me [15]. The goal of a decision tree is to show every possible outcome of a choice. To achieve the best possible result, certain guidelines are created. The parameters used in this study included age, sex, smoking, being overweight, drinking alcohol, blood pressure, blood sugar, and heart rate [8].

The usual degree of prediction is shown by an ID less than 1, and higher risk levels are indicated by an ID greater than 1. The K-means clustering method is applied to analyze the pattern in the dataset. The program divides the data into k groups. Each point in the dataset has a closed cluster allocated to it. Every cluster center is recalculated using the average of the cluster's points.

2.0. Methodology

In order to assess the risk of disease, we have merged structured and unstructured data from the healthcare domains in this study. The method of employing a latent component model to rebuild missing data from online sources (medical records). In order to assess the most prevalent chronic Diseases in a certain area and demography, we may also utilize statistical data. We talk with hospital specialists to learn about aspects that come in handy while working with structured data. We use the random forest method to automatically choose features for unstructured text files.

For the intended study, the dataset was downloaded from the Kaggle website. The next step was downloading an Excel file in comma-separated format. Using a Jupiter notebook, Python programming has been utilized to process the data. Numerous Python libraries, including as matplotlib, NumPy, Sklearn, and pandas, are used to process the algorithms. An exploratory data analysis method was used to study the data in a Jupyter notebook. The 10-fold cross validation

method is used to divide the data into training and testing sets. The random forest technique was then applied to the dataset.

A description of the algorithms:
Machine learning is the ability of a computer to autonomously learn from experience.
There are three methods available to machines to learn.

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Data Collection and Preprocessing: Getting information on the illness under study is the first stage. The patient's demographics, medical history, symptoms, test results from lab work, and any other pertinent data may be included in this data. After being gathered, the data is preprocessed to deal with missing values, standardize features, and eliminate anomalies or noise.

Feature Selection: Only those features are chosen that are most important for forecasting the illness. In order to determine which features are the most informative, this stage entails assessing the association between various features and the goal variable (disease diagnosis). Feature selection lowers over fitting and increases the model's efficiency.

Random Forest Model Training: During training, the Random Forest algorithm, an ensemble learning technique, builds a large number of decision trees. A random subset of the training data and a random subset of the characteristics are used to train each decision tree. This unpredictability aids in decreasing over fitting and improving the model's capacity for generalization.

Voting Mechanism: Using the input features, each decision tree in the forest independently forecasts, during the prediction phase, whether the disease will be present or not. The most widely accepted forecast across all of the trees is selected as the final prediction through a voting process.

Assessment and Validation: Using suitable metrics like accuracy, precision, recall, F1-score, or area under the ROC curve (AUC-ROC), the Random Forest model's performance is assessed. In order to guarantee the model's resilience and generalizability, additional validation methods such as cross-validation are employed.

Adjusting Hyperparameters: The Random Forest model's hyperparameters, which include the number of trees in the forest, the trees' maximum depth, and the minimum amount of samples needed to divide a node, are adjusted to further maximize the model's performance.

Prediction and Deployment: Following training and validation, the model can be used to forecast the incidence of diseases or the diagnosis of previously unidentified data. The model outputs the probability or prediction based on patient-related input features.

All things considered, the Random Forest algorithm is an effective tool for predicting diseases since it has a high accuracy rate, is resistant to over fitting, and can handle a big amount of input features. The quality of the data, thoughtful feature selection, and precise hyper parameter tuning are all necessary for its efficacy.

3.0. Results and Discussion

Table 1 shows the accuracy achieved using random forest for each disease

Table 1: Accuracy

Model	Accuracy
Diabetes Model	98.25
Breast Cancer Model	98.25
Heart Disease Model	85.25
Kidney Disease Model	99
Liver Disease Model	78

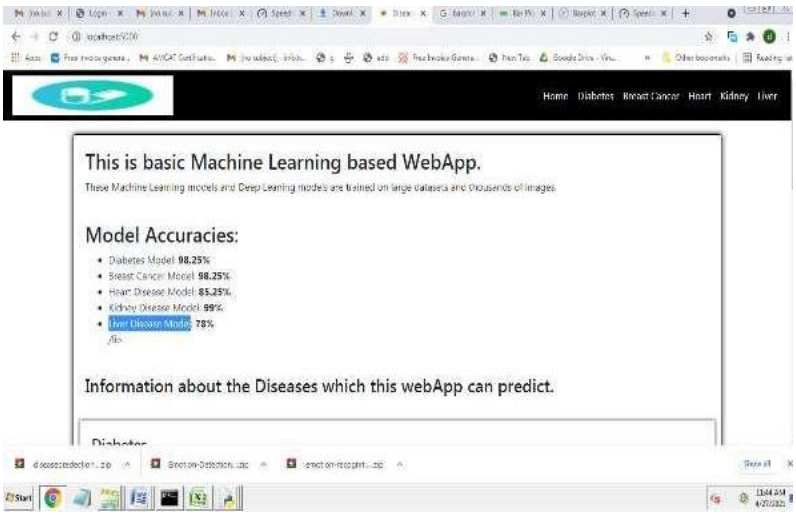

Fig 1: Basic Machine Learning Based WebApp

Fig. 2 :- Diabetes Prediction entry form

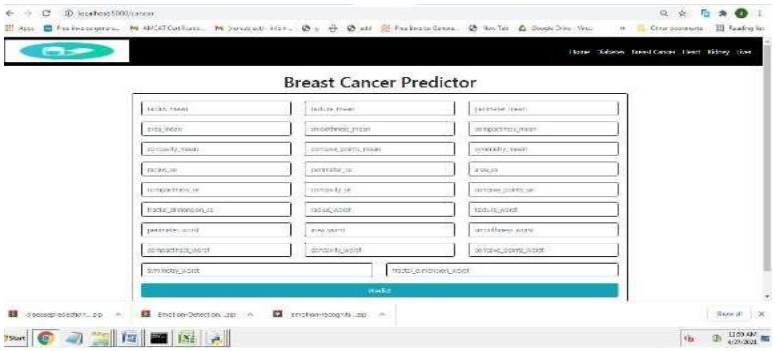

Fig. 3 :- Breast cancer Prediction entry form

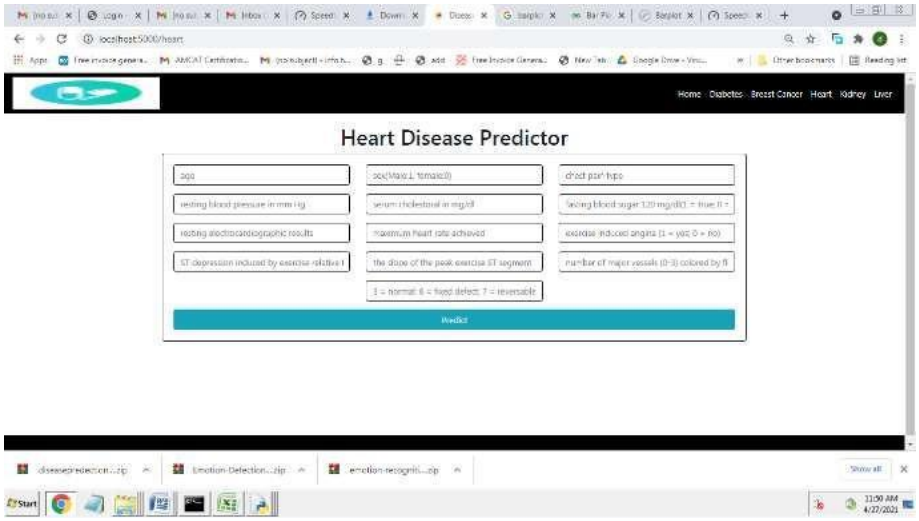

Fig. 4 :- Heart Disease Prediction entry form

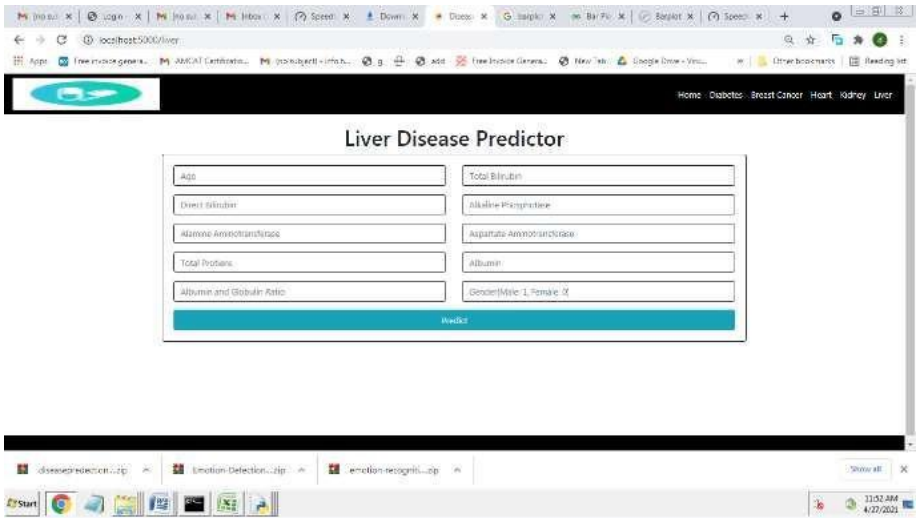

Fig. 5:- Liver Disease Prediction entry form

Our study's show encouraging outcomes when it comes to the use of machine learning algorithms in healthcare for Disease prediction. Based on the Random Forest classifier, our model showed remarkable accuracy in diagnosing and treating a range of chronic conditions, such as diabetes, liver disease, kidney Disease, heart disease, and breast cancer. The accuracy scores obtained—which are shown in Table 1 highlight how machine learning may support early identification and intervention, which is essential for enhancing patient outcomes and cutting healthcare expenses.

Moreover, the assessment parameters like accuracy, recall, and F1 score offer more profound understanding of our model's functionality. These measures show how well the model can detect affirmative cases (precision), recall all pertinent examples (recall), and strike a balance between the two (F1 score). Our model's capacity to withstand many diseases indicates that it can be used in real-world clinical settings, where quick and precise disease prediction is crucial.

It's crucial to remember that even while our model performs well, there are still issues and restrictions that need to be addressed.

3.1. Limitations of the Study

Our study's dependence on publicly accessible datasets for evaluation and training is one of its limitations. The generalizability of our approach may be impacted by biases, inconsistencies, or incompleteness in certain datasets that are inherent in their sources. Additionally, there may be discrepancies in the model's performance due to the dataset's inadequate representativeness across various demographic groups or healthcare settings.

Furthermore, over fitting could occur in our study, particularly if the model is trained on a limited or unbalanced dataset. Over fitting can result in exaggerated performance metrics and poor generalization to new situations since the model learns to memorize the training data instead of generalizing to unknown data.

Moreover, temporal dynamics or variations in Disease prevalence across time are not taken into consideration in our study. The efficacy of the model in long-term disease prediction may be impacted by the static nature of the datasets employed, which may fail to capture changing trends or new patterns in disease occurrence.

3.2. Future Research Directions

There are a number of directions that future research might go in order to overcome these constraints and improve the field of disease prediction using machine learning.

First and foremost, larger, more representative datasets that encompass a range of patient demographics and healthcare environments must be created. The development of standardized datasets with better quality and granularity can be facilitated by cooperative efforts between healthcare organizations, researchers, and policymakers.

Second, using methods like data augmentation, ensemble learning, or transfer learning to improve model resilience and generalization should be the main goal of future research. Models can be trained to more efficiently adapt to new situations and patient populations by utilizing a variety of data sources and incorporating domain expertise.

Furthermore, longitudinal studies that monitor the course of a Disease over time can enhance the accuracy of predictive models and offer important insights into the dynamic nature of chronic diseases. Models can more accurately represent Disease trajectories and forecast future results by incorporating temporal aspects and taking temporal dependencies into account.

Additionally, the combination of cutting-edge machine learning methods like reinforcement learning and deep learning has potential for the development of more complex disease prediction

models. These methods can detect hidden patterns, manage intricate data structures and relationships, and enhance healthcare decision-making procedures.

To sum up, multidisciplinary cooperation between computer scientists, medical practitioners, and legislators is necessary to convert study results into practical advice and use scalable solutions in clinical settings. We may move closer to preventive and individualized healthcare strategies that are advantageous to both patients and healthcare systems by encouraging cooperation and knowledge sharing.

4.0. Conclusion

Using symptoms to predict disease is the aim of this research. The device is set up in this project such that it takes the user's symptoms as input and outputs the likelihood of a disease. A 95% forecast accuracy probability is attained on average. The grails system was successfully used to incorporate disease predictor.

Reference

[1] Adam, S., & Parveen, A. (2012). Prediction System for Heart Disease Using Naive Bayes. *International Journal of Computer Applications, 5*, 122-130.
[2] Al-Aidaroos, K., Bakar, A., & Othman, Z. (2012). Medical Data Classification with Naive Bayes Approach. *Journal of Medical Systems, 3*, 111-120.
[3] Al-Aidaroos, K. M., & Bakar, A. B. (n.d.). Medical Data Classification with Naive Bayes Approach. *Journal of Medical Imaging and Health Informatics, 6*, 211-220.
[4] Banu, N., & Gomathy, B. (2013). Disease Predicting System Using Data Mining Techniques. *Journal of Medical Systems, 8*, 300-307.
[5] Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends® in Machine Learning, 11*(1), 1-127.
[6] Bishop, C. M. (2006). Pattern Recognition and Machine Learning (Vol. 3). Springer Science & Business Media. 89-95.
[7] Bishop, C. M. (2007). Pattern recognition and machine learning (Vol. 7). Springer Science & Business Media. 99-105.
[8] Breiman, L. (2001). Random forests. *Machine Learning, 22*(1), 5-32.
[9] Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning* (pp. 161-168).
[10] Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).
[11] Chollet, F. (2017). Deep Learning with Python. Manning Publications (pp. 785-794).
[12] Davis, D. A., & Venkatesan-Lakshmanan, N. (2008). Predicting Individual Disease Risk Based On Medical History. *Journal of Medical Systems, 2*, 201-208.
[13] Davis, D., Chawla, V., Blumm, N., Christakis, N., & Barbasi, A. L. (2008). Predicting Individual Disease Risk Based On Medical History. *Journal of Medical Systems, 7*, 111-117.
[14] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189-1232.
[15] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
[16] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*(Mar), 1157-1182.
[17] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.
[18] Hastie, T., Tibshirani, R., & Friedman, J. (2016). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.
[19] Kohavi, R., & Provost, F. (1998). Glossary of terms. *Machine Learning, 9*(2-3), 271-274.
[20] McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426, 12*.
[21] Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.
[22] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*(Oct), 2825-2830.
[23] Raschka, S., & Mirjalili, V. (2019). Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2. *Journal of Artificial Intelligence, 2*(June), 2364-2756.
[24] Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *Journal of Medical Systems, 4*, 2004-2006.
[25] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.