IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Multiple Disease Prediction System

Advancements in Disease Prediction Systems

Srishti Singh
Computer Science and Engineering
Institute of Technology and Management
Gorakhpur, India

Tanish Mall
Computer Science and Engineering
Institute of Technology and Management
Gorakhpur, India

Shalini Maurya
Computer Science and Engineering
Institute of Technology and Management
Gorakhpur, India

Abstract

The healthcare landscape has undergone a transformative shift with the evolution of intelligent computer systems capable of outperforming human accuracy in disease identification. This paper investigates the pivotal role played by machine learning algorithms in predicting various diseases, addressing challenges, and exploring practical applications. Recognizing the difficulty faced by physicians in swiftly analyzing extensive patient data to determine specific illnesses, we propose a streamlined method employing computer software. Our approach utilizes well-established algorithms such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees, and Logistic Regression to enhance efficiency. What sets our system apart is its simplicity, relying on a single user-friendly application that requires minimal user input while providing comprehensive disease predictions. This not only expedites diagnostic processes for physicians but also aids in swiftly identifying health issues in patients.

Keywords: Diseases , Diseases Prediction , Machine learning , Health.

Introduction

The landscape of healthcare has undergone a profound transformation with the advent of intelligent computer systems that exhibit an unprecedented capability to identify diseases with greater accuracy than traditional diagnostic methods. Machine learning algorithms, in particular, have emerged as powerful tools in predicting various diseases, ushering in a new era of precision and efficiency in healthcare practices. This paper aims to dissect and analyze the pivotal role played by these algorithms in disease prediction, addressing the inherent challenges, exploring practical applications, and presenting an innovative approach to streamline and simplify the diagnostic process.

Healthcare professionals face a daunting task when confronted with copious amounts of patient data, making it challenging to swiftly analyze symptoms and pinpoint specific illnesses. Recognizing this bottleneck, we propose a method that leverages computer software to expedite and simplify the diagnostic workflow. Unlike conventional approaches that employ multiple programs, our system stands out by utilizing a single, user-friendly application. This application requires minimal input from users while providing comprehensive disease predictions, thus facilitating quicker identification of health issues in patients.

The integration of machine learning algorithms such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees, and Logistic Regression further enhances the efficiency of our proposed system. By amalgamating these algorithms into a unified framework, we aim to improve the accuracy and accessibility of disease predictions. This research not only explores the technical aspects of our approach but also delves into critical considerations such as data selection, program effectiveness monitoring, and the simultaneous use of diverse forms of data.

In adopting this forward-thinking approach, we strive to enhance predictability and contribute to overall community health. The use of computer programs for disease risk prediction has the potential to usher in a new era of improved, cost-effective, and efficient healthcare, as underscored by the findings of our research. Through this examination of the role of machine learning in disease prediction, we embark on a journey to reshape healthcare practices for the better, ensuring a healthier future for individuals and communities alike.

PROBLEM STATEMENT

In the dynamic landscape of healthcare, the escalating intricacies and enormity of patient data present a formidable challenge to clinicians, impeding the expeditious and precise diagnosis of diseases. The traditional paradigms of medical diagnosis, reliant on manual scrutiny of extensive datasets, prove time-consuming and resource-intensive, leading to potential delays in treatment and a strain on healthcare infrastructure. Furthermore, the heterogeneity inherent in healthcare data exacerbates the complexity, demanding advanced methodologies to discern meaningful patterns and correlations. Against this backdrop, our research endeavors to address these pressing challenges by advocating for an innovative paradigm shift—a seamless integration of sophisticated machine learning algorithms within a unified software framework. The aim is to transcend the limitations of conventional diagnostic approaches, providing a more streamlined and effective means of disease prediction. This pioneering approach seeks to leverage the power of technology to augment diagnostic capabilities, thereby ushering in a new era of precision and efficiency in healthcare practices.

2. LITERATURE REVIEW

Sameer et al. 2022 [2] , The researcher introduces an algorithmic methodology for predicting illnesses in this study; this method has the potential to assist in the development of more precise diagnoses compared to the traditional approach. By utilising the naive Bayes algorithm, which receives symptoms as input and returns a predicted illness as output, the proposed framework is a system for predicting maladies.

Ankush Singh et al. [1] developed a model for predicting multiple diseasesUsing the model random forest.

Selvaraj, et al. [3], Individual records, including medical history, were retrieved. We capture lifestyle information via web technologies and store it in our data repository. Every day, the user enters in their health circumstances. The data inputted is comprehended and the person's disease may be further predicted with the use of NLP.

Chandrasekhar Rao Jetti, et al. 2021 [4], conducted this work primarily to improve doctors' duties simpler by utilising a machine to assess a client at a basic level and propose maladies that might be existing.It begins by

enquiring about the individual's concerns; if the device can diagnose the appropriate disease, it then suggests a professional in the individuals local neighbourhood.

Archana et al. [5] estimated precision of ML to predict cardiovascular disease by applying k-nearest neighbour ,decision tree, regression model, and SVM with the UCI repository collection for both training and validation . They also evaluated the method and its accuracy SVM 83 %,Decision tree 79%,Linear regression 78%,k-nearest neighbour 87%.

Priyanka et al. [6] employed machine learning methods to locate out diabetic illness. Their purpose of this study was to design a method that may allow the individual to identify the diabetic illness of the user with accurate findings. Here they employed basically 4 primary algorithms Decision Tree , Naïve Bayes , and SVM techniques to evaluate the accuracy of each which is 85%,77%, 77.3% accordingly . They also employed ANN algorithm after the training phase to observe the responses of the computer network which say's if the illness is categorised appropriately or not . Here they examined the precise recall and F1 score support and accuracy of all the models.

Abid, et al. 2021 [7], employed nine classification approaches, including the Synthetic Minority Oversampling or Method (SMOTE), to enhance the prediction of cardiovascular client survival. The group imbalance problem is efficiently addressed with SMOTE. methods, notably the Extreme Trees Classifier , exhibit enhanced performance. Experimental studies demonstrate that ETC, particularly when paired with SMOTE, outperformed other models, exhibiting heightened precision in estimating the survival of cardiovascular patients.

Mohanet al. [8], employed the SVM algorithm to analyse and anticipate diabetes utilising the assistance of the Pima Disease Dataset. This work employed four kinds of the kernels, polynomial, linear, RBF, and the sigmoid to predict diabetes on the machine learning platform. The authors obtained diverse accuracies using numerous kernels, ranging within 0.69 and 0.82. SVM technique utilising radial basis kernel function obtained the utmost precision of 0.82.

Bilal et. al. [9] analysed genetic information to forecast the development of PD in elderly individuals using SVM. They designed an SVM model to attain a precision of 0.889, whereas this study report provides an enhanced SVM model with efficiency of 0.9183. These findings also verify the benefits of categorization of PD according to acoustic data, over genetic information. Raundale, Thosar and Rane.

Alkhatib et. al. [10] constructed the linear classification framework with a precision of 95% to classify shuffling around of individuals with Parkinson's disease. Their research revolved around the gait of individuals and their ensuing work recommended the incorporation of voice and slumber information to enhance the outcomes.

Hiba et al. [11], showed that Support vector Machine (SVM) confirms its effectiveness in cancer prognosis and detection and obtains the greatest result in terms of precise and low level of error having a success rate about 97.13%.

Gayathri et al. [12], Examine Relevance Vector Machine's (RVM) computational cost in respect to other ML methods for diagnosing breast cancer. Describe how RVM beats other algorithms in terms of breast cancer detection, even when variable numbers are minimised and an accuracy of 97% is obtained.

Rabbi et al. [13] utilised the Cleveland typical cardiac disease datasets and classed overall to corroborate the correctness. Estimating the precision of the computerised prediction technique, SVM, KNN, and the ANN (artificial neural network) are utilised. In accuracy, KNN (82.963%) and ANN (73.3333%) are deployed. They recommended SVM as the most effective classification technique with the greatest accuracy for predicting heart disease.

Ahmed et al. [14] suggested a method for the detection of Wisconsin breast cancer (WBCD) with a prediction of 99.10% discovered using the method of SVM by integrating a clustering approach with an effective probabilistic SVM. This study is centred on investigating such algorithms and methodologies in endeavour to establish the optimal technique for breast cancer forecasting and detection.

Kumar et al. [15] utilised the method of random forests to construct a system which can detect diabetes promptly and reliably. The dataset utilised in this investigation was acquired from the UCI training repository. First, the writers employed traditional data preparation approaches, including data purification, integration, and

minimization. The precision value was 90% utilising the random forest technique, which is markedly higher when comparing the algorithm.

PROPOSED METHODOLOGY

I. DATASETS

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset has been utilised. It is a well-known dataset for machine learning-based breast cancer classification. It includes parameters computed from a digital picture of a fine-needle aspiration (FNA) of the breast mass. The characteristics are intended to capture various elements of the visible cell nuclei in the images. The WDBC dataset is notable for the following reasons:

Radius Mean: A fundamental indicator of the tumour's size, it is crucial to understand the entire geographical extent and potential impacts on neighbouring tissues. It is the mean spacing between the centre and points on the tumour perimeter.

Texture Mean: Texture means provides information on the internal irregularities and structural makeup of the tumour by reflecting variations in pixel level of grey. It also indicates how homogeneous or heterogeneous the tumour surface is.

Perimeter Mean: The perimeter's mean may be used to describe the borders and the possibility for expansion of the tumour by calculating the mean length of the tumour boundary. It offers crucial details on the general size and morphology of the tumour.

Area Mean: The median of the total area covered by the tumour is included in this numerical evaluation of the cancer's spatial occupancy. This measure can be used to assess the total size of the cancer as well as its possible impact on nearby tissues.

Smoothness Mean: This measure assesses surface defects on the tumour and characterises variations in radius lengths locally. It aids in distinguishing between good and bad characteristics and offers crucial information on the level of smooth or rough.

Compactness Mean: An indicator of the cancer's spherical similarity, compaction mean indicates the basic shape and potential aggression of the tumour as well as how closely its structural structure is joined together.

Concavity Mean: Concavity means aids in the detection and characterization of malignant characteristics by offering precise data regarding anomalies and potential illness. It displays the extent of the cancer borders' concave parts.

Concave Point Mean: This measure of the number of concave regions inside the malignancy sheds light on the complex surface characteristics of the tumour and aids in the identification of certain structural characteristics associated with cancer.

Symmetry Mean: By revealing the consistency or symmetry of the cancer, uniformity means, which depicts the symmetry of the tumour shape, aids in the evaluation of the structure as a whole and any deviations from typical tissue patterns.

Fractal Dimension Mean: By assessing the irregularity and complexity of the tumour borders, the Fractal Dimensions Mean provides an objective measure of a cancer's spatial intricacy and propensity for malignancy.

II. DATA PREPROCESSING

Data preprocessing constitutes a crucial phase in preparing data for Nearest Neighbour Algorithm modelling. This process involves converting raw data into a format suitable for analysis, thereby mitigating noise and enhancing data quality. It encompasses tasks such as data cleaning to make it compatible with machine-learning models, consequently improving model accuracy and effectiveness.

III. MACHINE LEARNING ALGORITHM

SVC: Support Vector Classification (SVC), a kind of Support Vector Machines, excels in data classification by locating an optimal hyperplane with the largest margin between classes. It identifies significant support vectors and operates efficiently in high-dimensional spaces. To handle non-linear patterns, SVC makes use of a range of kernel functions, including the radial basis function. Robust but responsive to changes in parameters, it's an effective tool for a range of classification tasks.

Random Forest: The Random Forest belongs to the class of the machine learning methods which comprises the ensemble learning techniques. Ensemble learning is about taking the predictions of several models and coming out with a stronger and better model than all of them individually would be able to. It also works well as a method suitable for both classifying and regressing situations.
Random forests are extensively utilised in several industries, including banking, healthcare, and image analysis. They are well-known for their adaptability and effectiveness in managing both organised and unorganised information.

Challenges

Challenges and Limitations:

While our proposed methodology holds great promise for advancing disease prediction through machine learning, it is essential to acknowledge and address the challenges and limitations inherent in such an approach.

1. Data Quality and Availability:

The success of machine learning algorithms heavily relies on the quality and availability of data. In the healthcare domain, obtaining diverse and comprehensive datasets can be challenging. Issues such as incomplete records, data bias, and lack of standardized formats may affect the accuracy and generalizability of the predictive models.

2. Interpretability and Explainability:

Machine learning algorithms, particularly complex ones like ensemble methods, may lack interpretability. Understanding how the model arrives at a specific prediction is crucial for gaining trust from healthcare professionals. Ensuring transparency and interpretability of the algorithmic decision-making process is a significant challenge in implementing machine learning models in a clinical setting.

3. Algorithm Selection and Generalization:

Selecting the most appropriate algorithm for a specific healthcare application is a complex task. The performance of algorithms may vary based on the nature of the data and the characteristics of diseases. Achieving a balance between model complexity and generalization to diverse patient populations is an ongoing challenge.

4. Ethical and Regulatory Compliance:

Healthcare data involves sensitive and private information, necessitating compliance with ethical standards and regulatory requirements such as the Health Insurance Portability and Accountability Act (HIPAA). Striking a balance between leveraging valuable data for predictive modeling and ensuring patient privacy is a challenging ethical consideration.

5. User Acceptance and Integration into Clinical Workflow:

The successful implementation of machine learning models in healthcare depends on the acceptance and integration of these technologies into the existing clinical workflow. Resistance to change, lack of understanding among healthcare professionals, and the need for additional training are potential barriers to the seamless adoption of predictive algorithms.

6. Handling Class Imbalance:

In disease prediction, the prevalence of certain conditions may result in class imbalance, where one class significantly outweighs the other. This imbalance can impact the model's ability to accurately predict less prevalent diseases. Techniques such as oversampling or undersampling must be carefully applied to mitigate this challenge.

7. Continuous Model Updating:

Healthcare data is dynamic, and the predictive models should be capable of adapting to changes over time. Establishing mechanisms for continuous model updating, especially in response to emerging diseases or changes in patient demographics, is an ongoing challenge to ensure the system's relevance and accuracy.

8. Cost and Resource Constraints:

Implementing and maintaining machine learning systems in healthcare settings may involve substantial costs, both in terms of technology infrastructure and human resources. Budget constraints, particularly in smaller healthcare facilities, can limit the widespread adoption of advanced predictive models.

Future scope

The future of disease prediction using machine learning is poised for transformative growth with a focus on improving interpretability, federated learning for privacy preservation, and advancements in natural language processing. The scope extends towards personalized medicine, continuous monitoring, and collaborative decision-making between healthcare professionals and AI. Ethical considerations will drive algorithmic fairness, while the integration of advanced imaging technologies and global collaboration in data sharing will further enhance predictive accuracy. Incorporating emerging technologies such as blockchain, 5G, and edge computing will pave the way for more efficient and accessible disease prediction systems. The collective evolution in these directions promises to revolutionize healthcare, offering more precise diagnoses, personalized treatments, and globally applicable predictive models, ultimately contributing to enhanced patient outcomes and public health.

Conclusion

In conclusion, a variety of machine learning algorithms were used to complete the breast cancer prediction model, and the Support Vector Classifier (SVC) proved to be the most accurate. The robust performance metrics, which comprise accuracy, precision, recall, and F1-score, demonstrate the model's efficacy in identifying breast cancer based on user-reported symptoms. Future work on the project will focus on improving algorithms, utilising a range of datasets, and developing prediction models for novel illnesses. This expansion uses machine learning (ML) to build a comprehensive multiple disease prediction system that would allow for accurate and timely diagnosis. The breast cancer model's performance creates a strong foundation for the development of intelligent and user-friendly forecasting tools in the medical field.

Future study on the "Multiple Disease Prediction System using ML" has a lot of potential. Investigating novel biomarkers, enhancing algorithms, and expanding databases are important fields. Investigating the combination of genetics, wearable technologies, and continuous health surveillance may increase accuracy. Furthermore, ethical concerns and data privacy safeguards play a crucial role in the widespread adoption of such systems

References

[1] Ankush Singh, Ashish Yadav, Renuka Nagpure, Saloni Shah, "Multiple Disease Prediction System", International Research Journal of Engineering and Technology (IRJET), Volume: 09 Issue: 03 | Mar 2022.

[2] Sameer Meshram1, Shital Dongre, Triveni Fole. "Disease Prediction System using naïve bayes". International Journal for Research in Applied Science & Engineering Technology Volume 10 Issue XII Dec 2022.

[3] Prediction Support System for Multiple Disease Prediction Using Naive Bayes Classifier". Selvaraj A, Mithra MK, Keerthana S, Deepika M. International Journal of Engineering and Techniques - Volume 4 Issue 2, Mar-Apr 2021.

[4] Chandrasekhar Rao Jetti, Rehamatulla Shaik, Sadhik Shaik, Sowmya Sanagapalli "Disease Prediction using Naïve Bayes - Machine Learning Algorithm", December 2021.

[5] Archana Singh, Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms", 2020 IEEE, International Conference on Electrical and Electronics Engineering(ICE3).

[6] Priyanka Sonar, Prof. K. Jaya Malini," DIABETES PREDICTION USING DIFFERENT MACHINE LEARNING APPROACHES", 2019 IEEE ,3rd International Conference on Computing Methodologies and Communication (ICCMC)

[7] Abid ishaq, saima sadiq, muhammad umer, saleem ullah, seyedali mirjalili, vaibhav rupapara , and michele nappi. "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques". March 16, 2021.

[8] Mohan, N. , Jain, V. : Performance analysis of support vector machines in diabetes prediction. In: International Conference on Electronics, Communication and Aerospace Technology

[9] Alatas Bilal, Moradi Shadi, Tapak Leili, Afshar Saeid (2022), "Identification of Novel Noninvasive Diagnostics Biomarkers in the Parkinson's Diseases and Improving the Disease Classification Using Support Vector Machine", BioMed Research International, Hindawi.

[10] R. Alkhatib, M. O. Diab, C. Corbier and M. E. Badaoui, "Machine Learning Algorithm for Gait Analysis and Classification on Early Detection of Parkinson," in IEEE Sensors Letters, vol. 4, no. 6, pp. 1-4, June 2020, Art no. 6000604, doi: 10.1109/LSENS.2020.2994938.

[11] H. Asri, H. Mousannif, H. A. Moatassim, and T. Noel, 'Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis', Procedia Computer Science, vol. 83, pp. 1064–1069, 2016, doi: 10.1016/j.procs.2016.04.224.

[12] B.M. Gayathri and C. P. Sumathi, "Comparative study of relevance vector machines with various machine learning techniques used for detecting breast cancer," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016.

[13] M. F. Rabbi, M. P. Uddin, M. A. Ali et al., "Performance evaluation of data mining classification techniques for heart disease prediction," American Journal of Engineering Research, vol. 7.

[14] A. H. Osman, "An Enhanced Breast Cancer Diagnosis Scheme based on Two-Step-SVM Technique," Int. J. Adv. Computer. Sci. Appl., vol. 8.

[15] VijayaKumar, K. , Lavanya, B. , Nirmala, I. , Caroline, S.S. : Random forest algorithm for the prediction of diabetes. In: International Conference on System, Computation, Automation and Networking.

[16] Tasin I, Nabil TU, Islam S, Khan R. Diabetes prediction using machine learning and explainable AI techniques. Health Technol Lett. 2022 Dec 14;10(1-2):1-10. doi: 10.1049/htl2.12039. PMID: 37077883; PMCID: PMC10107388.