Multiple Disease Prediction System Using Machine Learning

Harshit Gupta, Lakshay Dahiya
harshitgupta3801@gmail.com, lakshay07dahiya@gmail.com
Assistant Prof. (Dr.) Jyoti Kaushik (CSE Department)
Maharaja Agrasen Institute of Technology,
Rohini Sector-22, New Delhi, India

ABSTRACT

In the face of increasing health challenges, the "Multiple Disease Prediction System Using Machine Learning" project strives to address a critical concern: the proactive identification and prediction of various diseases for effective healthcare management. Fueled by the rising complexity of healthcare data and the need for personalized and timely interventions, this initiative aims to revolutionize disease prediction, prevention, and management through the integration of advanced machine learning techniques.

The motivation behind this project stems from the growing burden of multiple diseases and the imperative to shift from reactive to proactive healthcare strategies. Traditional healthcare models often fall short in anticipating and preventing diseases, leading to increased healthcare costs and compromised patient outcomes. By harnessing the power of machine learning algorithms, such as ensemble methods and deep learning models, the project seeks to analyze diverse health datasets, including genetic information, lifestyle factors, and historical medical records.

The project's scope is comprehensive, aiming to predict the likelihood of various diseases in individuals based on their unique health profiles. Machine learning models, particularly ensemble techniques like Random Forests or Gradient Boosting, serve as the core predictive

engines, leveraging diverse data sources to provide accurate and personalized risk assessments. The system integrates electronic health records, genetic data, and lifestyle inputs to offer tailored recommendations for disease prevention and early intervention.

Through an intuitive and user-friendly interface, the project envisions a centralized platform that empowers both healthcare providers and individuals to make informed decisions about health risks and preventive measures. The system will enable timely interventions, reducing the overall burden on healthcare systems and improving patient outcomes. By proactively identifying disease risks and promoting preventive measures, this initiative strives to usher in a future where healthcare is not only reactive but also predictive.

INTRODUCTION

In the ever-evolving landscape of healthcare, the integration of advanced technologies has become imperative for more accurate diagnostics and proactive health management. Among these technologies, machine learning stands out as a powerful tool capable of revolutionizing the way we approach disease prediction and prevention. This project, titled "Multiple Disease Prediction System Using Machine Learning," focuses on harnessing the potential of machine learning algorithms to predict and identify the risk of three prevalent diseases: diabetes, heart disease, and Parkinson's disease.

The contemporary healthcare system faces challenges in early detection and timely intervention for various diseases, often leading to severe health implications and increased healthcare costs. Traditional methods of disease prediction rely heavily on manual assessments and historical data, limiting their accuracy and efficiency. The incorporation of machine learning offers a paradigm shift by enabling the analysis of diverse datasets and the extraction of complex patterns, ultimately enhancing the predictive capabilities in healthcare.

Existing System

The existing disease prediction system lays the foundation for a comprehensive and efficient approach to addressing the challenges associated with predicting multiple diseases, including diabetes, heart disease, and Parkinson's disease. The system initiates with the crucial step of data collection, aiming to compile a vast and diverse dataset of medical records containing pertinent patient information and various medical features relevant to the target diseases. Subsequently, the collected data undergoes meticulous preprocessing to handle missing values, outliers, and ensure proper feature scaling, ensuring the dataset's quality and suitability for machine learning model training. The heart of the system lies in model training, where diverse machine learning algorithms, such as decision trees, random forests, and artificial neural networks, are employed to learn patterns and relationships within the preprocessed data. The system incorporates a robust model selection phase, utilizing performance metrics like accuracy, precision, and recall to identify the most effective algorithm for disease prediction. Rigorous model evaluation on an independent test dataset follows, providing insights into the selected model's accuracy and reliability in predicting multiple diseases. The final touch involves the development of a user-friendly interface tailored for healthcare professionals, facilitating seamless input of patient information and delivering predictions for multiple diseases, thereby culminating in a practical and accessible tool for disease prediction.

Proposed System

Addressing the challenge of predicting multiple diseases—diabetes, heart disease, and Parkinson's disease—requires a comprehensive approach integrating various aspects of data collection, model development, early detection, and community engagement. The following approach outlines a systematic plan to develop an effective Multiple Disease Prediction System using machine learning:

1. Data Collection: The first component of the system involves collecting a large dataset of medical records containing patient information and various medical features related to multiple diseases. This dataset will be used to train the machine learning models.

2. Data Preprocessing: The collected data will be preprocessed to handle missing values, outliers, and to perform feature scaling. This component of the system involves cleaning and preparing the data for model training.

3. Model Training: This component involves training different machine learning algorithms such as decision trees, random forests, and artificial neural networks on the preprocessed data. The trained models will be used for disease prediction.

4. Model Selection: The performance of different machine learning algorithms will be compared using metrics such as accuracy, precision, and recall, and the best- performing model will be selected for disease prediction.

5. Model Evaluation: The selected model will be evaluated on a separate test dataset to measure its accuracy and reliability in predicting multiple diseases. This component of the system involves testing the model and measuring its performance.

6. User Interface Development: The final component of the system involves developing a user-friendly interface that allows healthcare professionals to input patient information and receive predictions for multiple diseases. The interface will be designed to provide an easy-to-use tool for disease prediction.

Algorithm

LOGISTIC REGRESSION

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name logistic regression is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name multinomial logistic regression is usually reserved for the case when the dependent variable has three or more unique values, such as Married, Single, Divorced, or Widowed. Although the type of data used for the dependent variable is different from that of multiple regression, the practical use of the procedure is similar. Logistic regression competes with discriminant analysis as a method for analyzing categorical-response variables. Many statisticians feel that logistic regression is more versatile and better suited for modeling most situations than is discriminant analysis. This is because logistic regression does not assume that the independent variables are normally distributed, as discriminant analysis does.

This program computes binary logistic regression and multinomial logistic regression on both numeric and categorical independent variables. It reports on the regression equation as well as the goodness of fit, odds ratios, confidence limits, likelihood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform an independent variable subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values and provides ROC curves to help determine the best cutoff point for classification. It allows you to validate your results by automatically classifying rows that are not used during the analysis.

RANDOM FOREST

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

The first algorithm for random decision forests was created in 1995 by Tin Kam Ho[1] using the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg.

SVM

In classification tasks a discriminant machine learning technique aims at finding, based on an independent and identically distributed (iid) training dataset, a discriminant function that can correctly predict labels for newly acquired instances. Unlike generative machine learning approaches, which require computations of conditional probability distributions, a discriminant classification function takes a data point x and assigns it to one of the different classes that are a part of the classification task. Less powerful than generative approaches, which are mostly used when prediction involves outlier detection, discriminant approaches require fewer computational resources and less training data, especially for a multidimensional feature space and when only posterior probabilities are needed. From a geometric perspective, learning a classifier is equivalent to finding the equation for a multidimensional surface that best separates the different classes in the feature space.

SVM is a discriminant technique, and, because it solves the convex optimization problem analytically, it always returns the same optimal hyperplane parameter—in contrast to genetic algorithms (GAs) or perceptrons, both of which are widely used for classification in machine learning. For perceptrons, solutions are highly dependent on the initialization and termination criteria. For a specific kernel that transforms the data from the input space to the feature space, training returns uniquely defined SVM model parameters for a given training set, whereas the perceptron and GA classifier models are different each time training is initialized. The aim of GAs and perceptrons is only to minimize error during training, which will translate into several hyperplanes' meeting this requirement.

Conclusion

While extensive research has been dedicated to classifying skin conditions in humans, the study of skin disorders affecting animals, particularly lumpy skin disease in cows, has been relatively limited. This study addresses this gap by introducing a system that effectively distinguishes between animals with lumpy skin disease and those with normal skin, employing image processing and machine learning approaches. The developed technique exhibits a high degree of accuracy in identifying lumpy skin conditions, offering a promising avenue for efficient disease detection in livestock.

Future Enhancements

Several potential enhancements can further elevate the capabilities of the lumpy disease detection system. Firstly, the integration of Internet of Things (IoT) devices, such as cameras or sensors, presents an opportunity for real-time data capture from cows. This would enable continuous monitoring, enhancing the system's responsiveness to the dynamic nature of lumpy skin disease occurrences.

Additionally, the development of a mobile application could provide a user-friendly platform for individuals to upload images for analysis, receiving real-time results and facilitating convenient disease monitoring. Integration with existing electronic health record systems in veterinary clinics or farms is another valuable enhancement, streamlining data management and ensuring seamless incorporation into established workflows.

To keep pace with evolving challenges, continuous updates and refinements to the classification models are essential. This involves incorporating new data and leveraging advanced deep learning techniques to ensure the system's adaptability to emerging patterns and variations in lumpy skin disease.

In summary, these future enhancements aim to enhance the functionality, accessibility, and accuracy of the lumpy disease detection system. By incorporating IoT devices, mobile applications, integration with existing systems, advanced classification models, and continuous updates, the system can provide a comprehensive and efficient approach to disease monitoring and control measures for lumpy skin disease in cows.

References

1. H. EL Massari, S. Mhammedi, Z. Sabouri, and N. Gherabi, “Ontology-Based Machine Learning to Predict Diabetes Patients,” in Advances in Information, Communication and Cybersecurity, Cham, 2022, pp. 437–445. doi: 10.1007/978- 3-030-91738-8_40.

2. F. Alaa Khaleel and A. M. Al-Bakry, “Diagnosis of diabetes using machine learning algorithms,” Mater. Today Proc., Jul. 2021, doi: 10.1016/j.matpr.2021.07.196.

3. . J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," ICT Express, vol. 7, no. 4, pp. 432–439, Dec. 2021, doi: 10.1016/j.icte.2021.02.004.

4. P. Cıhan and H. Coşkun, "Performance Comparison of Machine Learning Models for Diabetes Prediction," in 2021 29th Signal Processing and Communications Applications Conference (SIU), Jun. 2021, pp. 1–4. doi: 10.1109/SIU53274.2021.9477824.

5. S. Barik, S. Mohanty, S. Mohanty, and D. Singh, "Analysis of Prediction Accuracy of Diabetes Using Classifier and Hybrid Machine Learning Techniques," in Intelligent and Cloud Computing, Singapore, 2021, pp. 399–409. doi: 10.1007/978-981-15-6202-0_41.

6. Aditi Gavhane, Gouthami Kokkula, Isha Panday, Prof. Kailash Devadkar, "Prediction of Heart Disease using Machine Learning", Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology(ICECA), 2021.

7. Md. Redone Hassan et al, "A Knowledge Base Data Mining based on Parkinson's Disease" International Conference on System Modelling & Advancement in Research Trends, 2019.

8. Santhana Krishnan J and Geetha S, "Prediction of Heart Disease using Machine Learning Algorithms" ICIICT, 2019.

9. Senthil kumar mohan, chandrasegar thirumalai and Gautam Srivastva, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" IEEE Access 2019.

10. M. Nikhil Kumar, K. V. S. Koushik, K. Deepak, "Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools" International Journal of Scientific Research in Computer Science, Engineering and Information Technology ,IJSRCSEIT 2019.

11. Dr. Anupam Bhatia and Raunak Sulekh, "Predictive Model for Parkinson's Disease through Naive Bayes Classification" International Journal of Computer Science & Communication vol. 9, Dec. 2017, pp. 194- 202, Sept 2017 - March 2018.