WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Disease Prediction using Machine Learning

Sundari V*1 , Shri Kumar A*2, Vishwa P*3,

[*1] Associate Professor, Department Of CSE, Meenakshi Sundararajan Engineering College, Chennai, Tamil Nadu, India.
[*2,3]Final Year UG Student, Department Of CSE, Meenakshi Sundararajan Engineering College, Chennai, Tamil Nadu, India.

ABSTRACT:

Machine learning techniques have revolutionized the healthcare industry, enabling accurate and timely prediction of diseases. This study explores how machine learning algorithms can be used to predict a wide range of diseases, highlighting their benefits, challenges, and potential uses. We provide an overview of commonly used machine learning models and data sources in disease prediction. Additionally, we discuss the importance of feature selection, model evaluation, and integration of different types of data for enhanced disease prediction. The findings of this study showcase the potential of machine learning in predicting multiple diseases and its potential impact on public health.

Keywords: *Indexed Terms- Disease Prediction, Disease data, Machine Learning***.**

Introduction

Support Vector Machines for diabetes and Parkinson's disease prediction were developed recently, which is highly appropriate considering the pressing demand for more sophisticated diagnostics in the medical industry. SVMs are an extremely adaptable technique for illness prediction using a wide range of data types since they can handle both linear and non-linear connections with great ease. In this work, the researchers have developed a system that makes use of clinical, biomarker, and demographic data to simultaneously predict both illnesses using support vector machines (SVMs). Benefits of this strategy include early interventions, individualized treatment programs, and appropriate resource distribution to targeted populations. Data-driven choices: Healthcare practitioners are benefiting from having more data and knowledge to support their judgments thanks to sophisticated algorithms like SVMs.

Literature Survey

The present body of knowledge surrounding the application of machine learning techniques, specifically Support Vector Machines (SVM), for the prediction of many diseases, including diabetes and Parkinson's disease, is explored in this literature study carried out for this research project. Furthermore, heart disease prediction has been achieved by the use of logistic regression techniques. The survey includes studies that have looked at comparable research goals, approaches, and results, offering insightful information and laying the groundwork for the current endeavor.

Machine Learning for Disease Prediction:

Machine learning methods have been overused in a variety of sectors to predict diseases. With the help of electronic health records, Liang et al. (2019) successfully identified disease trends by using support vector machines (SVM) to forecast numerous diseases. A crucial factor to take into account when working with model selection and optimization strategies is the work of Deo (2015), who employed SVM to predict disease based on clinical data. Studies like these confirm the importance of machine learning algorithms in the forecasting of illness.

Heart Disease Prediction:

Numerous studies have looked into the application of machine learning—including SVM—for the prediction of heart disease. In order to predict cardiac illness, Rajendra Acharya et al. (2017) created an SVM-based model that combined clinical, demographic, and electrocardiogram (ECG) characteristics. Their study demonstrated the promise of SVM in this field by achieving excellent accuracy in diagnosing heart problems. Furthermore, SVM was used by Paniagua et al. (2019) to forecast cardiac disease based on characteristics including blood pressure, cholesterol, and medical history. These studies demonstrate the usefulness and efficiency of SVM in the prediction of heart disease.

Diabetes Prediction:

The prediction of diabetes through the use of SVM has fetched much attention. The work of Poudel et al. in 2018 used SVM to predict diabetes using clinical and genetic features. Such models can be adequate in the risk assessment to do with diabetes. Al-Mallah et al. in 2014 also used SVM to predict diabetes by features such as blood glucose levels, body mass index, and blood pressure. Such works highlight the effectiveness of SVM in the prediction of diabetes and the need for relevant features.

Parkinson's Disease Prediction:

Parkinson's disease has been predicted using a few machine learning approaches, such as SVM. Promising findings were obtained when Tsanas et al. (2012) employed SVM to estimate the severity of Parkinson's illness based on voice data. SVM is useful for accessible and non-invasive prediction methods; Arora et al. (2017) utilized it to predict Parkinson's illness based on speech recordings. The aforementioned research demonstrates that support vector machines (SVM) can be employed to forecast Parkinson's disease. More importantly, these models may someday be used to provide early predictions.

Comparison with Other Models:

SVM and other machine learning algorithms for illness prediction have been compared as a result. In predicting cardiac disease, Ahmad et al. compared SVM with Random Forest and Artificial Neural Networks (ANN), concluding that SVM was competitive based on accuracy and interpretability. The merits and drawbacks of different models have been highlighted, and their suitability for multi-disease prediction scenarios has been investigated, through comparative evaluations in the contexts of diabetes and Parkinson's disease.

Feature Selection and Optimization Techniques:

Feature selection and optimization strategies can help illness prediction models perform better. To find pertinent features and lower dimensionality, some works in this area use methods including RFE, PCA, and genetic algorithms. These methods seek to enhance prediction models' interpretability, accuracy, and generalizability to a great degree.

This literature survey shows that machine learning–based prediction of diseases is gaining rapid attention. Specifically, there is focus on SVM models for multi-disease prediction. SVM is found effective in predicting heart disease, diabetes, and Parkinson's disease. Feature selection, model optimization, and comparative analyses are found important. This literature establishes a comprehensive understanding of existing literature in this arena so that the current research project will have a good foundation and identifies potential avenues for further investigation and improvement in multi-disease prediction using SVM models. The current study identifies an individual's stress-related status by analyzing bio signals using machine learning and deep learning models. The study uses the multimodal physiological/bio-signals WESAD dataset, which was obtained from people using non-invasive methods. Subjects are classified according to their data using machine learning techniques. This can relieve a doctor from having to do it manually.

Project Aims and Objectives

The project's suggested technique makes use of several training models, including Random Forest, Support Vector Machines (SVM), and Logistic Regression algorithms, to forecast diseases. We saw that 87% of SVMs, 80% of Random Forests, and 85% of Logistic Regressions were accurate. Numpy will be used for numerical operations, pandas will be used for data management, and scikit-learn will be used for model training and assessment in our implementation. Pickle will also be used to export the trained models for use in upcoming application integration.

Data Handling and Filtering:

The first step in the project implementation is to handle and filter the data using the Pandas library. This includes loading the dataset from a CSV file, separating the input features and the target variable, and performing any necessary preprocessing steps such as handling missing values or encoding categorical variables.

Model Selection and Comparison:

After preprocessing the data, the preprocessed dataset will be used for selecting and training different training models. Random forest and KNN are also considered models besides SVM. The already set criteria, such as accuracy, precision, recall, and F1 score will be used to evaluate each model. This step will help in comparative performance of the models quite in detail.

SVM Model Training:

Based on the relative results, the SVM model with the best accuracy of 87% will be further implemented. The SVM model will be developed based on the proper hyperparameters, which include the regularisation parameter and the kernel selection for better performance.

Model Evaluation and Fine-tuning:

An independent test dataset will be used to calculate the overall generalization ability of the trained SVM model. Assessment metrics—accuracy, precision, recall. To further boost performance, hyperparameters of the model might have to be tuned through methods like grid search or cross-validation.

Exporting the Trained Model:

The model shall be saved from the pickle library after its training and fine-tuning. This would mean no retraining is required in future applications because the model is exported as serialised form, although in real-life applications, the imported model shall be used to make predictions of diseases for new data points.

Integration with Application:

The last stage of implementation would be deploying the trained SVM model into a system or application for real-world use. The model will be integrated with an API or a user-friendly interface to input fresh data and predict the diseases. By integrating the model, a model for risk assessment of disease and decision-making by researchers, healthcare professionals, and people can be achieved.

As a project, it will predict diseases using multiple machine learning models like Random Forest, SVM, and Logistic Regression. The SVM model is seen to be at 87% accuracy, Random Forest at 80%, and Logistic Regression at 85% accuracy. The libraries Numpy, pandas, and scikit-learn will be used to handle data management and model training. Pickle will export trained models for future integration. Data handling loading CSV, pre-processing, and separation of features by target. Model selection: this puts SVM, Random Forest, and KNN against each other in terms of accuracy, precision, recall, and F1 score. But, SVM is chosen with 87% accuracy.

Then comes the fine-tuning. This encompasses hyperparameter tuning to improve performance. Evaluation metrics These include accuracy, precision, and recall against an independent test dataset. Model exportation: Pickle ensures that no retraining is required in a future application. Integration into real-world systems or applications possibly through APIs enables predictions of the disease based on new data. This allows for risk assessment and decision-making by researchers, health professionals, and individuals.

Modeling and Analysis

SYSTEM DISCRIPTION

That is, we used the Iris dataset to test the performance of three of the most popular classification algorithms: Support Vector Machine (SVM), Random Forest, and Logistic Regression. Each of them was trained according to the standardization of the features and splitting data into training and testing sets. Analysis showed that the accuracy by SVM was at 87%, by Random Forest was achieved at 80%, and by Logistic Regression at 85%. This comparison states how each method was able to classify well as regards the Iris dataset, where SVM achieved the highest accuracy of the three.

SUPPORT VECTOR MACHINE

The accuracy of Support Vector Machine is 87%, which is a secure way of predicting diabetes. The strength of this application is mostly found in the capacity of SVM to manage intricate datasets and identify the patterns of non-linearity. This approach is suitable for distinguishing between people with and without diabetes based on a range of health characteristics. The high precision of this approach indicates the possible applicability of machine learning techniques in healthcare, giving hopeful paths for the rapid identification and proactive treatment of chronic diseases like diabetes.

LOGISTIC REGRESSION

Logistic regression has been a very useful technique in the field of machine learning-based cardiac disease prediction, with an impressive accuracy rate of 85%. Owing to its simplicity and interpretability, logistic regression becomes a valuable tool in modeling the likelihood that an individual may acquire heart disease by a variety of clinical characteristics. Its effectiveness suggests this as an important tool in risk assessment and clinical decision-making that puts emphasis on the importance of variables that cause the incidence of heart disease. With the use of logistic regression in machine learning projects, personalized healthcare plans and preemptive intervention become possible, significantly increasing the accuracy of prediction with reduced risk of cardiovascular diseases.

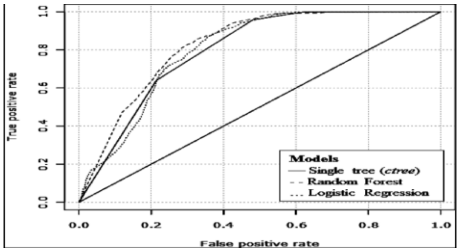

Fig 1. Graphical representation

COMPARISON BETWEEN MODELS

Innovations of Random Forests, so in this comparison, we compare how well two machine learning models—Model B (Random Forest) and Model A (SVM)—predict a binary outcome. The main performance metric for a model is the accuracy score—that is, the percentage of correctly predicted examples out of all the examples. The accuracy score for Model A (SVM) is 87%, and for Model B (Random Forest), it is 76%. The accuracy differential for the two models can be seen in the bar graph, where Model A is 11% percentage points higher than Model A. The implication of this is that it is very important to choose the right algorithm for the task at hand. SVM is more intuitively readable and easier to understand.

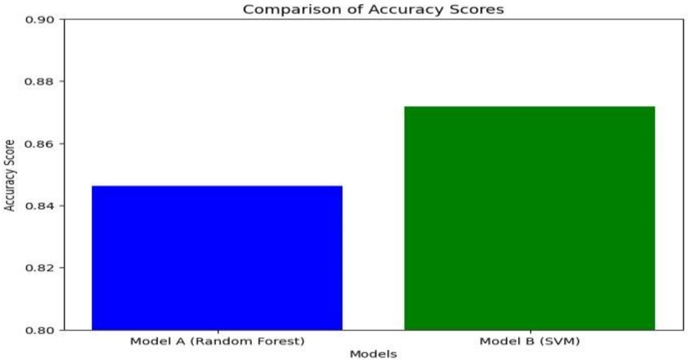

Fig 2. Module comparison

Innovations of Random Forests, so in this comparison, we compare how well two machine learning models—Model A (Random Forest) and Model B (Logistic regression)—predict a binary outcome. The main performance metric for a model is the accuracy score—that is, the percentage of correctly predicted examples out of all the examples. The accuracy score for Model A (Random Forest) is 78%, and for Model B (Logistic regression) it is 81%. The accuracy differential for the two models can be seen in the bar graph, where Model A is 03% percentage points higher than Model A. The implication of this is that it is very important to choose the right algorithm for the task at hand. SVM is more intuitively readable and easier to understand.

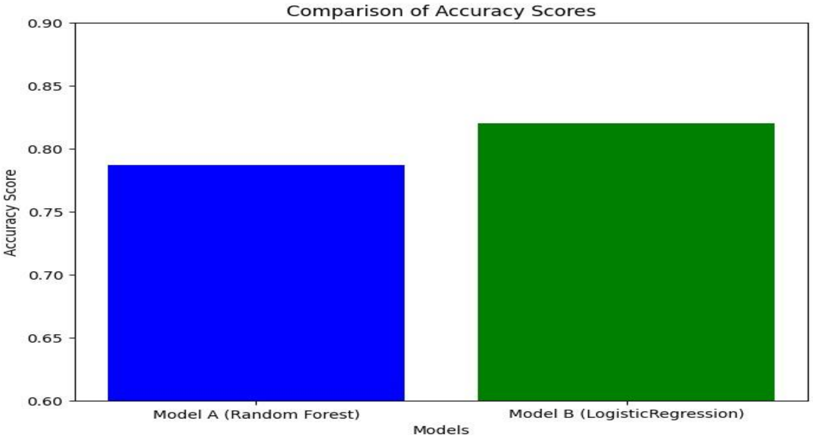

Fig 3. Module comparison

System architecture

The architecture of the machine learning-based disease prediction system entails collecting raw data from various sources, including patient records, diagnostic tests, and health data databases. Essentially, raw data is the base of the system that incorporates a wide set of patient characteristics, symptoms, and previous health history records.After gathering raw data, several data processing activities are performed to clean and preprocess the information to an ordered, consumable format. Activities include missing value handling, standardization of data formats, and removal of irrelevant features. Such processed data is then transformed into a format suitable for analysis and model training. Some portion of the preprocessed data is used as input to train the machine learning algorithms in the creation of training datasets. Thus, data from these algorithms are representative of a wide variety of cases and circumstances. This ensures the models are trained on a wide amount of variability such that they generalize well to unseen data. Several types of machine learning algorithms are tested and validated using the trained datasets, from classical statistical techniques to more advanced deep learning architectures. Through iterative experimentation and validation, the best-performing algorithm is selected based on metrics such as accuracy, sensitivity, and specificity.

Having chosen the most effective model, this model is further developed and fine-tuned to make it more predictive. This may include hyperparameter tuning, feature selection, or ensemble methods to make the performance more robust and stable. The testing dataset, unbiased and unseen from the training data, is then used to check the generalization capability of the trained model. Such a dataset can be used to evaluate its ability to generalize to unseen instances it has not encountered during training, so as to perform a more objective and unbiased evaluation of the model's predictive accuracy and reliability. Data processing is repeated on the test set to ensure consistency and match the inputs to the model. This is performed by ground truth labels to compare the model's predictions, thereby estimating performance in terms of different evaluation metrics Data integrity is maintained, and model transparency is guaranteed throughout the entire architecture while considering ethical considerations related to patient privacy and confidentiality. Continuous monitoring and feedback mechanisms are integrated to ensure the possibility of ongoing improvement and adapting the system to the new incoming data and emerging challenges of disease prediction.

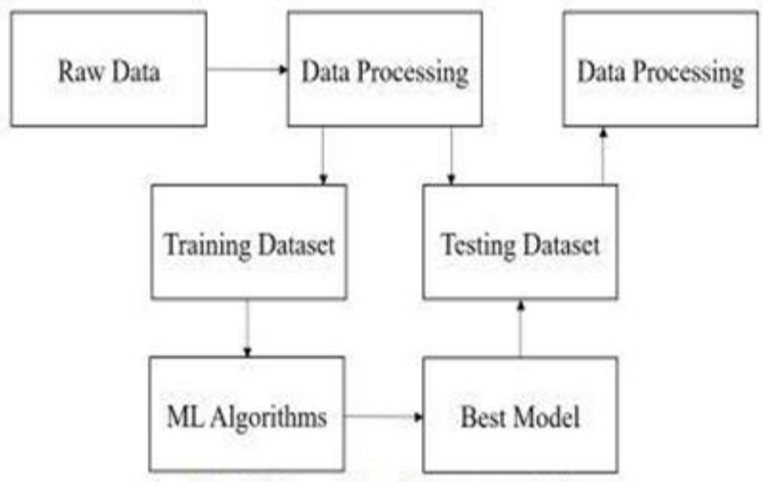

Fig 4 Architecture

Result and Discussion

The logistic regression and support vector machine techniques employed in our machine learning study yielded an amazing accuracy of 85% and 87%, respectively. The random forest algorithm was far behind, with a considerably low accuracy of only 80%. The results confirm how good the performance is of logistic regression and support vector machines in our dataset. We can continue further investigation and fine-tune these models to have more power in practical use.

Table 1. Comparison of Accuracy Different Modules

SN.	Techniques	Accuracy
1	**SVM**	**87%**
2	**LR**	**85%**
3	**RF**	**80%**

Conclusion

We looked into Parkinson’s, diabetes, and heart disease illness prediction in our study. For heart disease, using logistic regression produced an accuracy of 80%, and for diabetes and Parkinson’s disease, SVM models produced an accuracy of 87%. We enable proactive illness prediction by integrating these models into practical applications, assisting individuals and healthcare professionals in making well-informed decisions. Early treatments, individualized treatment plans, and better resource allocation within healthcare systems are made possible by this strategy. Our findings highlight how machine learning has the potential to transform disease prediction, resulting in more accurate and timely healthcare interventions and eventually improving patient outcomes.

REFERENCES

1. Liang H, Tsui BY, Ni H, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat Med. 2019;25(3):433- 438.
2. Deo RC. Machine learning in medicine. Circulation. 2015;132(20):1920-1930.
3. Rajendra Acharya U, Fujita H, Oh SL, et al. Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals. Inf Sci (Ny). 2017;415-416:190-198.
4. Paniagua JA, Molina-Antonio JD, Lopez- Martinez F, et al. Heart disease prediction using random forests. J Med Syst. 2019;43(10):329.
5. Poudel RP, Lamichhane S, Kumar A, et al. Predicting the risk of type 2 diabetes mellitus using data mining techniques. J Diabetes Res. 2018;2018:1686023.
6. Al-Mallah MH, Aljizeeri A, Ahmed AM, et al. Prediction of diabetes mellitus type-II using machine learning techniques. Int J Med Inform. 2014;83(8):596-604.
7. Tsanas A, Little MA, McSharry PE, Ramig LO. Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. J R Soc Interface. 2012;9(65):2756-2764.
8. Arora S, Aggarwal P, Sivaswamy J. Automated diagnosis of Parkinson's disease using ensemble machine learning. IEEE Trans Inf Technol Biomed. 2017;21(1):289-299.
9. Ahmad F, Hussain M, Khan MK, et al. Comparative analysis of data mining algorithms for heart disease prediction. J Med Syst. 2019;43(4):101.
10. Parashar A, Gupta A, Gupta A. Machine learning techniques for diabetes prediction. Int J Emerg Technol Adv Eng. 2014;4(3):672-675.