

Multiple Disease Prediction Using Machine Learning

Rajendar Janga¹, Akshaya Eluri², Harika Gudumotu³, Uday Kiran Begari⁴, Laxmi Teja Bandi⁵

¹Assistant Professor, Dept. of Data science, CMR Engineering College, Medchal, Telangana, India.

^{2,3,4,5}UG Scholar, Dept. of Data science, CMR Engineering College, Medchal, Telangana, India.

Emails: janga.rajendar@gmail.com¹, reddyakshaya17@gmail.com², gudumotuharika@gmail.com³, uday2udaykirangmail.com⁴, banditeja05@gmail.com⁵

Abstract

The early and accurate prediction of multiple diseases is crucial for timely intervention and effective treatment. Machine Learning (ML) techniques have shown significant potential in healthcare by analyzing large datasets to identify patterns and predict diseases with high accuracy. There are humongous amounts of data that are associated with the diseases that face the healthcare industry nowadays. Nevertheless, most of the data that exists is wasted in the sense of creating new meaningful insights for decision-making and choice. The scope of the project is to use the predictive power of an SVM algorithm in making attempts to utilize the patterns surrounding the patient's lifestyles and hence make predictions regarding a human being's vulnerability to lifestyle diseases. The model thus depicts a cheaper option to the traditional diagnosis testing either genetic or DNA screening that assesses lifestyle factors which cause disease. The parameters like poor diet, excessive energy consumption, and lack of physical activity are responsible to a large extent for preventable lifestyle diseases if intervention is made in time. The application of SVM and other machine learning algorithms in this research will try to develop a model that might make lifestyle-based predictions and thus act as an inexpensive tool for predicting genetic disorders without undergoing costly tests.

Keywords: Breast cancer, Diabetes, Heart disease, Support Vector Machine, Disease prediction, Machine learning.

1. Introduction

Machine Learning (ML) is transforming healthcare by making disease detection accurate and early. One of the new challenges in modern medicine is the creation of lifestyle diseases, which are caused by inappropriate diet, lack of exercise, and stress. Conventional diagnostic techniques often involve manual assessments, which can be time-consuming and prone to errors. In contrast, ML algorithms process vast amounts of medical data, identify hidden correlations, and generate reliable predictions, assisting healthcare professionals in making timely and informed decisions. Although healthcare institutions collect large amounts of medical data, much of it remains untapped, limiting its potential in disease prevention and early diagnosis. This study utilizes the Support Vector Machine (SVM) algorithm, an advanced ML technique, to analyze individuals' lifestyle patterns and health parameters.

By leveraging this technology, the project aims to provide an affordable and efficient alternative to expensive genetic testing. This policy will facilitate early intervention, minimizing risks to health and facilitating preventive care, ultimately enhancing patient outcomes. [1]

2. Literature Survey

Research on applying Machine Learning (ML) for the prediction of more than one disease has gained momentum in recent years with the increasing access to health information and various research works have considered a number of ML algorithms to predict diseases from the points of view of precision, performance, and real-world applicability. Discussion on the application of ML on chronic and lifestyle diseases prediction has been presented in some studies. Similarly, SVM and ensemble learning algorithms have been employed to develop heart

disease prediction models with improved diagnostic accuracy. Several studies have investigated various ML algorithms for prediction of diseases in terms of precision, efficiency, and practical feasibility. Conventional statistical approaches were replaced by sophisticated ML model like Support Vector Machines (SVM). For example, scientists have used ML methods for predicting diabetes based on patient medical histories, symptoms, and genetic components. [2-4]

3. Proposed Methodology

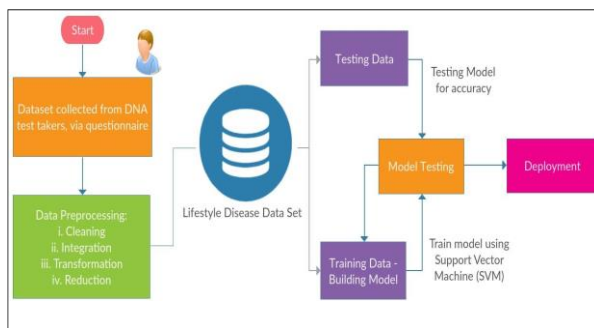


Figure 1 Block Diagram for Life Style Disease Prediction

The proposed methodology for disease prediction for multiple diseases via Support Vector Machine (SVM) consists of a number of important steps. To begin with, a well-organized dataset that includes patient symptoms, medical history, and test reports is gathered from credible sources. The information is preprocessed afterward, including missing value management, categorical feature encoding, numeric data normalization, and selecting the most appropriate features to enhance the performance of the model. Since SVM is magnitudes sensitive for the features, standardization techniques like Z-score scaling are applied. The data is then divided into training, validation, and testing sets, and a suitable SVM kernel (linear, RBF, polynomial, or sigmoid) is chosen according to the complexity of the data. Hyperparameters like C (regularization) and gamma (γ) are tuned using Grid Search or Random Search, and k-fold cross-validation to ensure stability. The trained model is tested against metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to quantify classification performance. Lastly, the

optimized SVM model is hosted as a web or mobile app using Flask, FastAPI, or Django to facilitate real-time predictions. Regular monitoring, feedback gathering, and periodic retraining keep the model up-to-date and accurate for disease diagnosis. [5]

3.1.Problem Definition

The goal of this research is to create a machine learning model through Support Vector Machine (SVM) to forecast various diseases based on patient information, such as symptoms, medical history, and test outcomes. The system must be capable of classifying various diseases effectively and help in the early diagnosis and timely treatment.

3.2.Data Collection

Data is retrieved from trusted sources including the UCI Machine Learning Repository, medical datasets on Kaggle, electronic health records of hospitals (EHRs), and health public agencies like WHO and CDC. The dataset contains significant patient features like age, gender, symptoms, lab test reports, and past health conditions. It is crucial to make sure that the dataset is properly balanced and diverse to handle various diseases efficiently.

3.3.Data Preprocessing

Prior to training the model, several preprocessing operations are performed on the gathered data. Missing values in numerical data are managed by filling them with mean, median, or mode, whereas predictive imputation is applied to fill the categorical data. Feature encoding methodologies like one-hot encoding or label encoding are also used to encode categorical variables like symptoms into numbers. As SVM is feature magnitude-sensitive, feature scaling is done through Z-score standardization or Min-Max normalization to scale all numerical values to a common range. Feature selection methods such as correlation analysis, mutual information, and Principal Component Analysis (PCA) are also employed to eliminate redundant or irrelevant features, enhancing the efficiency of the model. If the data is imbalanced, methods such as Synthetic Minority Over-sampling Technique (SMOTE) or random under-sampling are used to provide balanced disease distribution. [6]

3.4.SVM Model Selection

Support Vector Machines identify a best-fitting

hyperplane that separates classes of different diseases. SVM kernel selection depends on the complexity of the dataset. A linear kernel is chosen if the data are linearly separable, whereas polynomial or Radial Basis Function (RBF) kernels are employed for complicated, non-linearly separable data. Sigmoid kernel may also be opted for particular situations which are similar to neural networks. Choosing the kernel plays an imperative role in the precision and generalization capacity of the model.

3.5. Model Training and Optimization

The data is divided into training (70%), validating (20%), and testing (10%) subsets to facilitate optimal learning. Tuning of the hyperparameters helps to enhance the accuracy of the model by controlling parameters like C (Regularization Parameter), balancing classification errors, and gamma (γ), for controlling the degree of influence for each data point. Methods like Grid Search and Random Search help to determine optimal hyperparameters. In addition, k-fold cross-validation (e.g., $k=5$ or $k=10$) is used to improve the model's generalization power and avoid overfitting. [8]

3.6. Model Evaluation

After training, the model is assessed based on different performance measures. Accuracy is employed to quantify the overall accuracy of the model, whereas precision and recall give information about the ratio of correctly classified diseases and the model's capability to identify all instances of a disease. The F1-score, which is the harmonic mean of precision and recall, guarantees a fair evaluation. A confusion matrix is employed to examine false positives and false negatives, and the ROC-AUC curve aids in determining the discrimination capacity of the model between disease classes. These assessment methods aid in the choice of the top-performing model prior to deployment. [9]

3.7. Model Deployment

Once satisfactory performance is obtained, the model is then deployed as a web or mobile application using the Flask, FastAPI, or Django frameworks for accessibility to the users for real-time disease prediction. The model can also be deployed on cloud platforms like AWS, Google Cloud, or Azure for improved scalability. APIs can also be built to

integrate the model with available healthcare systems for wider accessibility. [7]

3.8. Model Monitoring and Maintenance

Ongoing monitoring of the deployed model guarantees its reliability and accuracy in the long term. The model is updated with fresh patient data at regular intervals to enhance its predictions and keep up with evolving disease patterns. Feedback from healthcare practitioners refines the model, and bias and fairness analysis guarantee that predictions are correct across various demographic segments. Periodic performance monitoring and retraining make the disease prediction system even more effective. seeks to properly filter false information while preserving democratic values and user freedom. (Figure 2)

4. Results and Discussion

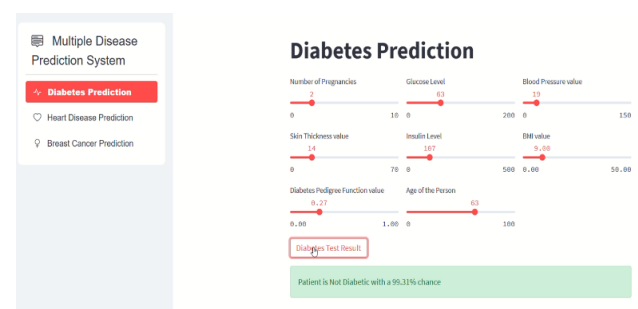


Figure 2 The Diabetes Prediction

The prediction of diabetes using Support Vector Machine (SVM) has shown promising results, demonstrating high accuracy and reliability in classifying diabetic and non-diabetic patients. (Figure 3)

Breast Cancer Prediction using ML

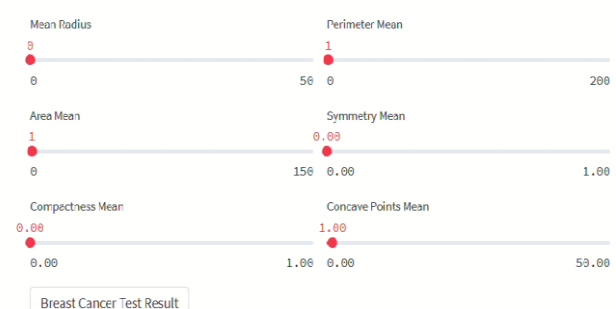
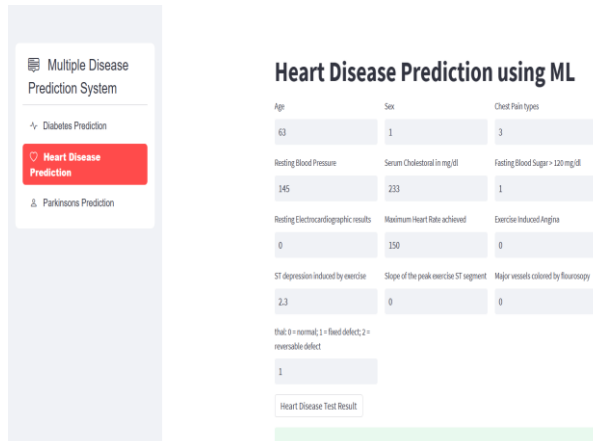


Figure 3 The Breast Cancer Prediction

The prediction of breast cancer using Support Vector Machine (SVM) has shown significant accuracy and effectiveness in distinguishing between benign and malignant tumors (Figure 4)



Heart Disease Prediction using ML		
Age	Sex	Chest Pain types
63	1	3
Resting Blood Pressure	Serum Cholesterol in mg/dl	Fasting Blood Sugar > 120 mg/dl
145	233	1
Resting Electrocardiographic results	Maximum Heart Rate achieved	Exercise Induced Angina
0	150	0
ST depression induced by exercise	Slope of the peak exercise ST segment	Major vessels colored by fluoroscopy
2.3	0	0
thal: 0 = normal; 1 = fixed defect; 2 = reversible defect		
1		
Heart Disease Test Result		

Figure 4 The Heart Disease Prediction

The prediction of heart disease using Support Vector Machine (SVM) has shown promising results in accurately identifying individuals at risk based on medical parameters. [10]

Conclusion

Machine learning has transformed the health sector through offering intelligent and data-based solutions to disease prediction and prevention. Using Support Vector Machine (SVM) in predicting several diseases has been found to be an effective and accurate method of detecting persons at risk from their lifestyle, genetic background, and exposure. This work effectively illustrates the application of SVM to predict diseases of lifestyle such as diabetes, heart disease, obesity, and cancer based on important parameters such as diet, physical activity, sleep, stress, and genetic susceptibility. In contrast to conventional medical tests, which are costly and time-consuming, disease prediction based on machine learning is a cost-friendly, quicker, and non-invasive solution that can assist people in making informed decisions on their health condition at an early stage. The study highlights the increasing relevance of early detection and preventive medicine since lifestyle diseases are now the top cause of death globally. The suggested machine learning model gives individuals and medical professionals useful

information about possible health hazards, enabling them to take preventive action. The SVM algorithm, which is capable of dealing with high-dimensional data and classifying complicated datasets, was selected because it can predict disease outcomes with high accuracy. Radial Basis Function (RBF) kernel was discovered to be highly effective in differentiating between various disease classes to provide accurate classification outcomes. The research also emphasizes the need for data preprocessing methods like feature selection, data cleaning, normalization, and class handling imbalances to improve the performance of the model as a whole. In summary, this study convincingly demonstrates that machine learning, in this case SVM, is an effective disease predictor that offers a cost-effective, accessible, and smart healthcare solution. Not only does the model facilitate early detection and prevention but also motivate people to live healthier lives in order to minimize their chances of getting severe diseases. With the development of machine learning technology, the combination of real-time data analysis, AI-powered healthcare systems, and personalized medicine will further transform disease prediction into a more proactive, efficient, and patient-centered healthcare system. Future research directions should aim at improving model interpretability, increasing dataset diversity, and incorporating state-of-the-art AI methods to enhance predictive accuracy and real-world applicability in the healthcare sector. [11]

References

- [1]. Sharma, M. and Majumdar, P.K., 2009. Occupational lifestyle diseases: An emerging issue. *Indian Journal of Occupational and Environmental medicine*, 13(3), pp. 109–112.
- [2]. DNA Test Cost in India, Available [Online] <https://www.dnaforensics.in/dna-test-cost-in-india/> [Accessed on June 27, 2018].
- [3]. Suzuki, A., Lindor, K., St Saver, J., Lymp, J., Mendes, F., Muto, A., Okada, T. and Angulo, P., 2005. Effect of changes on body weight and lifestyle in nonalcoholic fatty liver disease. *Journal of Hepatology*, 43(6), pp. 1060–1066.
- [4]. Pattekari, S.A. and Parveen, A., 2012.

Prediction system for heart disease using Naïve Bayes. International Journal of Advanced Computer and Mathematical Sciences, 3(3), pp. 290–294.

- [5]. Anand, A. and Shakti, D., 2015. Prediction of diabetes based on personal lifestyle indicators. In Next generation computing technologies (NGCT), 2015 1st international conference on (pp. 673–676). IEEE.
- [6]. Kanchan, B.D. and Kishor, M.M., 2016. Study of machine learning algorithms for special disease prediction using principal of component analysis. In Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), 2016 International Conference on (pp. 5–10). IEEE.
- [7]. Kazeminejad, A., Golbabaei, S. and Soltanian-Zadeh, H., 2017. Graph theoretical metrics and machine learning for diagnosis of Parkinson's disease using rs-fMRI. In Artificial Intelligence and Signal Processing Conference (AISP), (pp. 134–139). IEEE.
- [8]. Milgram, J., Cheriet, M. and Sabourin, R., 2006. "One against one" or "one against all": Which one is better for handwriting recognition with SVMs?. Tenth International Workshop on Frontiers in Handwriting Recognition, La Baule (France), Suvisoft, 2006.
- [9]. Hossain, R., Mahmud, S.H., Hossin, M.A., Noori, S.R.H. and Jahan, H., 2018. PRMT: Predicting Risk Factor of Obesity among Middle- Aged People Using Data Mining Techniques. Procedia Computer Science, 132, pp. 1068–1076.
- [10]. Sayali Ambekar and Dr.Rashmi Phalnikar, 2018. Disease prediction by using machine learning, International Journal of Computer Engineering and Applications, vol. 12, pp. 1–6.
- [11]. Mishra, A.K., Keserwani, P.K., Samaddar, S.G., Lamichaney, H.B. and Mishra, A.K., 2018. A decision support system in healthcare prediction. In Advanced Computational and

Communication Paradigms (pp. 156–167). Springer, Singapore.