Predicting Disease(s) using Machine Learning in Medical Science

Gurpreet Kaur¹, Gagandeep Jagdev²*,

¹Research Scholar, M.Tech (CE), Yadavindra College of Engineering, Talwandi Sabo. Punjab, India.
²Technical Officer, Punjabi University Guru Kashi Campus, Damdama Sahib, Talwandi Sabo, Punjab, India.

*Corresponding Author: Gagandeep Jagdev, Technical Officer, Punjabi University Guru Kashi Campus, Damdama Sahib, Talwandi Sabo, Punjab, India

Abstract: Machine learning techniques have gained significant attention in the field of disease prediction due to their ability to analyze complex healthcare data and provide accurate predictions. This research paper presents an investigation into the application of machine learning algorithms for disease prediction. The research findings highlight the potential of machine learning algorithms in predicting diseases and supporting clinical decision-making. The ability to identify high-risk individuals at an early stage can lead to timely interventions, personalized treatment plans, and improved patient outcomes. The research paper showcases various real-world applications of machine learning in disease prediction, spanning a wide range of medical conditions, including cancer, cardiovascular diseases, diabetes, and infectious diseases. The conducted research contributes to the growing body of knowledge on the application of machine learning in disease prediction. The results underscore the significance of utilizing advanced computational techniques in healthcare to harness the power of large-scale data analysis and enhance disease prediction capabilities. The findings have implications for healthcare providers, policymakers, and researchers in improving disease prevention, diagnosis, and treatment strategies.

Keywords: Accuracy, Diseases, Disease prediction, Disease symptoms, Machine Learning.

1. INTRODUCTION

Disease prediction models can identify the likelihood of an individual developing a specific disease based on their risk factors, symptoms, or biomarkers. Early detection allows for timely intervention, enabling healthcare professionals to implement preventive measures, initiate appropriate treatments, or recommend lifestyle modifications. Early intervention can lead to better health outcomes, reduced morbidity, and potentially even save lives [1]. Early disease prediction is vital for timely intervention, improved treatment outcomes, reduced disease burden, preventive measures, cost savings, improved quality of life, and effective public health planning. By identifying diseases at their early stages, healthcare systems can achieve better health outcomes, optimize resource allocation, and ultimately improve the well-being of individuals and communities [2].

Disease prediction at early stages is of utmost importance for several key reasons:

- **Timely Intervention:** Early disease prediction allows for timely intervention and treatment, which can prevent the progression of diseases or minimize their impact. By identifying diseases at their early stages, healthcare providers can initiate appropriate treatments, therapies, or lifestyle modifications that have a higher likelihood of success.

- **Improved Treatment Outcomes:** Early detection and intervention often lead to better treatment outcomes. Diseases detected early are generally at a more manageable and less advanced stage, increasing the chances of successful treatment and recovery. Early treatment can prevent complications, reduce disease severity, and improve long-term prognosis [3].

- **Minimized Disease Burden:** Early disease prediction can reduce the burden of disease on individuals, healthcare systems, and society as a whole. By detecting diseases early, healthcare resources can be utilized more efficiently, reducing the need for costly interventions, hospitalizations, and long-term management of advanced diseases.
Preventive Measures: Early disease prediction enables the implementation of preventive measures to reduce the risk of disease occurrence or progression. This may include lifestyle modifications, regular screenings, vaccinations, or targeted interventions aimed at addressing specific risk factors. Preventive strategies can be more effective when implemented before irreversible damage occurs [4].

Cost Savings: Early disease prediction can lead to significant cost savings in healthcare. By identifying diseases at their early stages, healthcare costs associated with advanced treatments, hospitalizations, and long-term management of chronic conditions can be minimized. Early intervention is often more cost-effective than managing complications or treating advanced-stage diseases.

Quality of Life: Early disease prediction can significantly improve the quality of life for individuals. By detecting and treating diseases early, individuals can experience fewer symptoms, maintain functional abilities, and have a better overall health status [5]. Early intervention can help individuals continue their daily activities and reduce the impact of diseases on their physical, emotional, and social well-being.

Reducing Disease Transmission: Early disease prediction is crucial for contagious diseases or outbreaks. Rapid detection and early intervention can help contain the spread of infectious diseases, limit transmission to others, and prevent epidemics or pandemics. Early identification of individuals at risk allows for appropriate isolation, contact tracing, and implementation of preventive measures [6].

Targeted Healthcare Planning: Early disease prediction provides valuable insights into disease patterns, risk factors, and population health trends. This information allows healthcare systems and public health authorities to plan and allocate resources more effectively, focusing on prevention, early detection, and targeted interventions for high-risk populations or regions [7].

The significance of disease prediction using machine learning is multifaceted and has far-reaching implications for individuals, healthcare providers, and public health. Some key points highlighting the significance of disease prediction:

- **Improved Patient Care:** Disease prediction empowers healthcare providers with valuable information about their patient's future health risks. This knowledge enables personalized and proactive care, as clinicians can tailor their approach to each patient's specific needs. Predictive models provide insights into the probability of disease occurrence, progression, or recurrence, allowing for individualized treatment plans and targeted interventions.

- **Resource Allocation and Cost Management:** Disease prediction can help healthcare systems allocate resources effectively. By identifying high-risk individuals or populations, healthcare providers can prioritize screening programs, diagnostic tests, or preventive interventions where they are most needed. This targeted allocation of resources optimizes healthcare delivery, improves cost-effectiveness, and reduces unnecessary procedures or treatments for low-risk individuals.

- **Public Health Planning:** Disease prediction models contribute to public health planning and resource allocation at a broader scale. By analyzing data from populations or regions, predictive models can assist in identifying disease hotspots, outbreaks, or trends. This information can guide public health officials in implementing preventive measures, resource distribution, and policy interventions to mitigate the impact of diseases on a larger scale [8].

- **Health Risk Communication and Patient Empowerment:** Disease prediction models facilitate health risk communication between healthcare providers and patients. By conveying personalized risk assessments and probabilities, patients can make informed decisions regarding their lifestyle choices, health behaviors, and adherence to preventive measures. Disease prediction empowers individuals to take an active role in their health management and engage in proactive measures to reduce their disease risk.
• Research and Drug Development: Disease prediction models provide valuable insights into the underlying mechanisms, risk factors, and biomarkers associated with diseases. This information aids researchers in identifying new targets for drug development, designing clinical trials, and understanding disease progression. Predictive models can also support precision medicine initiatives by identifying subgroups of patients who may respond differently to specific treatments [9].

• Public Health Surveillance and Early Warning Systems: Disease prediction models can be integrated into public health surveillance systems, monitoring data streams to detect early signals of disease outbreaks or epidemics. By analyzing patterns and anomalies in health data, these models can provide early warning signals, enabling public health officials to take proactive measures, allocate resources, and implement containment strategies promptly.

Overall, disease prediction using machine learning has the potential to revolutionize healthcare by enabling proactive, personalized interventions, optimizing resource allocation, and improving health outcomes at both individual and population levels. It contributes to a shift from reactive to proactive healthcare, with a focus on prevention, early detection, and targeted interventions, ultimately leading to better patient care and public health outcomes.

2. DISEASE PREDICTION TECHNIQUES USING MACHINE LEARNING CLASSIFIERS

Machine learning classifiers can predict diseases by analyzing a dataset of symptoms using a variety of techniques. The general overview of the process is mentioned as under.

• Data collection: A dataset is created by collecting information about symptoms and their corresponding diseases. This dataset should be labeled, meaning each instance (row) should have the symptoms and the corresponding disease or condition it represents.

• Data preprocessing: The collected data needs to be preprocessed to make it suitable for the machine learning classifier. This involves tasks such as removing irrelevant or duplicate data, handling missing values, and normalizing or scaling the data if necessary [10].

• Feature extraction: In this step, relevant features are extracted from the data on the symptoms. Feature extraction involves selecting or transforming the data into a format that captures the relevant information for the classification task. This can include techniques such as dimensionality reduction (e.g., principal component analysis) or extracting specific features from the data of the symptoms.

• Training the classifier: Once the data is preprocessed and features are extracted, the machine learning classifier is trained on the dataset. Various classification algorithms can be used, such as decision trees, random forests, support vector machines (SVM), or deep learning models like neural networks. During training, the classifier learns the patterns and relationships between symptoms and diseases in the dataset [11].

• Model evaluation: After the classifier is trained, it needs to be evaluated to assess its performance. This is typically done by using evaluation metrics such as accuracy, precision, recall, and F1-score. The dataset may be divided into training and testing subsets to evaluate the classifier's performance on unseen data [12].

• Predicting diseases: Once the classifier is trained and evaluated, it can be used to predict diseases based on new or unseen symptom data. The classifier takes the symptoms as input and applies the learned patterns to make predictions about the most likely disease or condition associated with those symptoms [13].

It's important to note that the effectiveness of a machine learning classifier in predicting diseases depends on the quality and representativeness of the dataset, the choice of features, and the algorithm used. Additionally, it's crucial to continuously update and refine the classifier as new data and information become available to improve its accuracy and reliability.

3. RESEARCH METHODOLOGY

The research work to be conducted can be divided into two phases mentioned under and shown in the flowchart in Fig. 1.
Create or obtain a dataset comprising multiple attributes specifying the symptoms of a particular disease. Perform training and testing of a dataset and conduct classification using different machine learning classifiers. Conduct classification to obtain the readings of performance evaluation metrics like accuracy, precision, recall, and F1-score.

Perform prediction of diseases using different ML classifiers.

**Fig 1. Flowchart of research methodology**

4. **Implementation And Results**

This section depicts the implementation phase of the anticipated research work. The proposed idea of constructing an interface for predicting diseases based on analyzing the symptoms has been implemented in Python. 8 different Machine Learning (ML) classifiers have been implemented. The dataset has been derived from the Kaggle data repository. The complete dataset consists of 2 CSV files. One of them is for training and the other is for testing the model. Each CSV file has 133 columns. 132 of these columns are symptoms that a person experiences and the last column is the prognosis. These symptoms are mapped to 42 diseases you can classify these sets of symptoms.

Five symptoms are selected from the list of symptoms given in the dataset. Each classifier performs its classification and reaches a decision and predicts the possible disease. The values of the performance evaluation parameters like accuracy, precision, recall, and F1 score are calculated with which the disease has been predicted. A GUI interface has been designed to offer a user-friendly interface to the users. Fig. 2 shows the programmed research work as a front-end interface to predict the diseases by analyzing the symptoms. The five symptoms are to be selected from the drop-down comprising a list of the symptoms.

**Fig 2. Designed Interface for predicting diseases**
Fig. 3 shows the name of the patient and five symptoms selected from the five dropdowns, one in each case.

The measure for calculating the four performance evaluation parameters is mentioned in Table 1 as under.

**Table 1. Performance Metrics**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>TP / (TP + FN)</td>
</tr>
<tr>
<td>Accuracy</td>
<td>(TP + TN) / (TP + FN + FP + TN)</td>
</tr>
<tr>
<td>Precision</td>
<td>TP / (TP + FP)</td>
</tr>
<tr>
<td>F1-score</td>
<td>2 * (Recall * Precision) / (Recall + Precision)</td>
</tr>
</tbody>
</table>

The diseases identified by different ML classifiers are shown in Fig. 4.
The diseases identified by different ML classifiers in Fig. 4 are listed in Table 2.

**Table 2. Predicted Diseases by different ML Classifiers**

<table>
<thead>
<tr>
<th>ML classifiers</th>
<th>Predicted Diseases</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT</td>
<td>Drug Reaction</td>
</tr>
<tr>
<td>RF</td>
<td>Drug Reaction</td>
</tr>
<tr>
<td>NB</td>
<td>Drug Reaction</td>
</tr>
<tr>
<td>GB</td>
<td>Drug Reaction</td>
</tr>
<tr>
<td>LR</td>
<td>Bronchial Asthma</td>
</tr>
<tr>
<td>KNN</td>
<td>Allergy</td>
</tr>
<tr>
<td>AB</td>
<td>GERD</td>
</tr>
<tr>
<td>SGD</td>
<td>Chicken Pox</td>
</tr>
</tbody>
</table>

Table 3 depicts the values of different performance evaluation parameters.

**Table 3. Values of performance metrics**

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F1 Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT</td>
<td>90.12195</td>
<td>89.64417</td>
<td>90.12195</td>
<td>89.13899</td>
</tr>
<tr>
<td>RF</td>
<td>91.95122</td>
<td>90.64868</td>
<td>91.95122</td>
<td>90.68549</td>
</tr>
<tr>
<td>NB</td>
<td>92.68293</td>
<td>91.23113</td>
<td>92.68297</td>
<td>91.34896</td>
</tr>
<tr>
<td>GB</td>
<td>91.82926</td>
<td>90.81222</td>
<td>91.82926</td>
<td>90.68710</td>
</tr>
<tr>
<td>LR</td>
<td>93.53658</td>
<td>91.17199</td>
<td>93.53658</td>
<td>91.93586</td>
</tr>
<tr>
<td>KNN</td>
<td>4.63414</td>
<td>0.39519</td>
<td>4.63414</td>
<td>0.71380</td>
</tr>
<tr>
<td>AB</td>
<td>63.90243</td>
<td>65.78777</td>
<td>63.90243</td>
<td>63.71491</td>
</tr>
<tr>
<td>SGD</td>
<td>91.70731</td>
<td>90.16736</td>
<td>91.70731</td>
<td>90.28806</td>
</tr>
</tbody>
</table>

Fig. 5 shows the graphical representation of the results illustrated in Table 3. LR has achieved the highest accuracy of 93.53658%, recall of 93.53658%, and F1 Score of 91.93586%. NB achieved the highest precision value of 91.23113% followed by LR with the precision value of 91.17199%. The lowest values have been obtained for the KNN classifier with an accuracy of 4.63414, precision of 0.39519, recall of 4.63414, and F1 Score of 0.71380. So, it can be concluded that LR outperformed other ML classifiers.

![Performance of ML Classifiers at 80-20%](image)

**Fig 5. Comparative graphical representation of the results obtained in Table 3**

So, it can be concluded that as per computed values of performance evaluation metrics, the best-predicted disease is Bronchial Asthma.

5. CONCLUSION

In conclusion, this research paper highlights the significance of disease prediction using machine learning classifiers in the field of medical science. Through a comprehensive review of approaches and applications, it becomes evident that machine learning holds great potential for improving early diagnosis, treatment outcomes, and public health planning. Real-world applications across various medical conditions demonstrate the effectiveness of machine learning in disease prediction. From cancer to cardiovascular diseases, diabetes, and infectious diseases, machine learning models have been successfully deployed for early detection, risk assessment, and personalized treatment planning.
The implementation of the proposed model for predicting diseases by analyzing the symptoms as per the methodology has been conducted and it is found that LR outperforms other ML classifiers in terms of performance. LR scored the highest accuracy, recall, and F1 score. The precision is best scored by NB which too is not far away from the one scored by LR. Furthermore, the role of interpretability is emphasized, as it plays a crucial role in the adoption of machine learning models by medical professionals. Efforts to develop explainable AI techniques are underway to provide insights into the decision-making process of complex models, facilitating their integration into clinical workflows.

REFERENCES


**AUTHORS’ BIOGRAPHY**

**Gurpreet Kaur**, is a Research Scholar pursuing M.Tech in Computer Engineering. She is working in the capacity of Principal at Institute of Technical Education and Industrial Training since 2012. Her research interest is Exploring Medical Science using Machine Learning and Deep Learning techniques.

**Dr. Gagandeep Jagdev**, is currently serving in the capacity of Technical Officer at Punjabi University Guru Kashi Campus, Damdama Sahib (PB). His total teaching and research experience is more than 16 years and has 158 International and National publications in reputed journals and conferences to his credit. He is the guest editor of the Journal of Imaging, MDPI. He is also a member of the editorial board of several reputed International Journals indexed in ESCI, Scopus, ACM, WoS, and Pubmeds and has been an active Technical Program Committee member of several International and National conferences conducted by renowned universities and academic institutions. He has been bestowed with Best Research Paper awards 4 times by NITTTR (Chandigarh), Government College of Engineering and Technology (Jammu), and Guru Nanak College, Budhlada (PB). He has actively participated in more than 80 Webinars and FDPs. His field of expertise is Image Processing, Big Data Analytics, Data Science, Cloud Computing, Cloud Security, Cryptography, and WANETs.