Feasible Prediction of Multiple Diseases using Machine Learning

Banoth Ramesh[1*], *G.* Srinivas[1], *P.* Ram Praneeth Reddy[1], *MD* Huraib Rasool[1], *Divya* Rawat[2], *Madhulita* Sundaray[3]

[1]Department of CSE (AI & ML), GRIET, Hyderabad, Telangana State, India
[2]Uttaranchal Institute of Technology, Uttaranchal University, Dehradun, 248007, India
[3]KG Reddy College of Engineering & Technology, Hyderabad, India

Abstract. Automated Multiple Disease Prediction System using Machine Learning is an advanced healthcare application that utilizes machine learning algorithms to accurately predict the likelihood of a patient having multiple diseases based on their medical history and symptoms. The system employs a comprehensive dataset of medical records and symptoms of various diseases, which are then analysed using machine learning techniques such as decision trees, support vector machines, and random forests. The system's predictions are highly accurate, and it can assist medical professionals in making more informed decisions and providing better treatment plans for patients. Ultimately, the viable Multiple Disease Prediction System using Machine Learning has the potential to improve healthcare outcomes and reduce healthcare costs by predicting and preventing disease early.

1 Introduction

Machine Learning (ML) gives computers the capability to learn without explicitly programmed. ML is the technology that no one come across. It is apparent from its name that makes the computer to work more similar to humans - the learning ability.

Arthur Samuel coined the term machine learning was coined in 1959. The term ‘machine learning’ discusses gathering information that may be measured or evaluated and then used to train a machine learning model. How effectively a machine-learning model works is significantly influenced by the quantity and quality of data utilized for training and testing. Data can take many different formats, including numerical, category, or time-series data, and can originate from a range of sources, including databases, spreadsheets, and APIs. Machine learning algorithms use data to find relationships and patterns between input parameters. Applications of ML are as follows:

1. Google Maps
2. Linkedin
3. Facebook
4. Medical Imaging
5. Art Restoration

* Corresponding author: ramesh1702@grietcollege.com

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

6. Forensic Science
7. Robotics
8. Satellite Imaging

The advantages of ML are as follows:

1. Ability to learn complex representations
2. End-to-end learning
3. Improved accuracy
4. Improved performance
5. Scalability and parallel processing
6. Cost computation

Developing a Machine learning implementation comes with its fair share of challenges and pitfalls. Firstly, acquiring and preparing a high-quality dataset can be time-consuming and expensive, especially for niche domains. Secondly, choosing the right architecture and hyper parameters requires extensive experimentation and can be computationally demanding. Thirdly, over fitting is a common pitfall, where the model performs well on training data but fails to generalize to unseen examples. Fourthly, training deep learning models often requires significant computational resources. Lastly, deploying and scaling learning models in production can be complex and may require additional infrastructure and maintenance efforts.

2 Existing methods

The automated Multiple Disease Prediction System using Machine Learning is an advanced healthcare application that utilizes machine learning algorithms to accurately predict the likelihood of a patient having multiple diseases based on their medical history and symptoms. The system employs a comprehensive dataset of medical records and symptoms of various diseases, which are then analysed using machine learning techniques such as decision trees, support vector machines, and random forests. The system's predictions are highly accurate, and it can assist medical professionals in making more informed decisions and providing better treatment plans for patients. Ultimately, the Multiple Disease Prediction System using Machine Learning has the potential to improve healthcare outcomes and reduce healthcare costs by predicting and preventing disease early.

Authors [16] highlighted the significance of ML in prediction, pattern recognition and error reduction across diverse fields, emphasizing the impact of AI in broad domain. Authors [17] presented text classification algorithms for various applications and explores the use of machine learning in detecting phishing attacks. Authors [18] discussed the use of machine learning and neural networks, especially CNN, for recognizing handwriting patterns, with a focus on Telugu film industry names, achieving high accuracy (98.3%). The paper [19] discussed the role of Intelligent Decision Support Systems (IDSS) in Healthcare Monitoring, especially for heart disease. Results claimed that IDSS enhances decision-making functionalities in uncertain healthcare scenarios, there by significantly improving the monitoring and remedial activities. Authors [20] suggested data mining techniques to predict disease-prevalence based on symptoms in healthcare data. The appropriate prediction helps healthcare organizations avoid drug shortages and further ensures timely treatment of patients. The paper [21] explores the distinct ML applications in predicting heart attacks using patient health records. It compares Random Forest and CNN methods, and findings showed that Random Forest’s better performance in terms of accuracy.

Table 1. Summary of existing approach.

Ref. No.	Dataset	Accuracy
[1]	Data Collected from Local Hospitals and Open Sources	Random Forest Preferred well compared to Naive Bayes and Decision tree
[2]	Dataset obtained from Cleaveland Heart Disease database at UCI Repository	Highest Accuracy achieved by KNN with 87%
[3]	Dataset Obtained from UCI Repository	Highest Accuracy achieved by SVM with 89.34%
[4]	Dataset from different patients experienced with different disease	CNN - MDRP algorithm performed much better for unstructured data
[5]	Dataset used is public clinical available dataset	In the model generated SVM gave high precision and accuracy
[6]	Dataset Obtained from UCI Repository	A study on algorithms shown that SVM and LR Algorithms performed well.
[7]	Dataset is obtained from University of California, Irvine	Highest Accuracy achieved by Logistic regression 82.89%
[8]	Dataset Obtained from UCI Repository	Highest Accuracy achieved by Naive Bayes 88.163%
[9]	Dataset is obtained from University of California, Irvine	Highest Accuracy achieved by Logistic regression 82.89%
[10]	Dataset used in this research is available at web source	Rapid Miner tool should higher degree of correctness than Matlab and Weka tool
[11]	Dataset Obtained from UCI Repository	Highest Accuracy shown by Naive Bayes 86.6%
[12]	Dataset Obtained from UCI Repository	Highest Accuracy shown by Naive Bayes 86.6%
[13]	Dataset Obtained from UCI Repository	The DT, LR and MLP algorithms showed maximum precision and minimum errors among all algorithms
[14]	Data collected from Hospitals, this data consists of both structured and unstructured data	Decision Tree, Regression model has the best prediction accuracy in the models generated
[15]	Datasets available in “Kaggle” and “University of California, Irvine (UCI) database”	In the model generated SVM gave high precision and accuracy

3 Problem statement and objectives

3.1 Problem Statement

The Multiple Disease Prediction System using Machine Learning is an advanced healthcare application that utilizes machine learning algorithms to accurately predict the likelihood of a patient having multiple diseases based on their medical history and symptoms. The system employs a comprehensive dataset of medical records and symptoms of various diseases, which are then analysed using machine learning techniques such as decision trees, support vector machines, and random forests. The system's predictions are highly accurate, and it can assist medical professionals in making more informed decisions and providing better treatment plans for patients. Ultimately, the Multiple Disease Prediction System

using Machine Learning has the potential to improve healthcare outcomes and reduce healthcare costs by predicting and preventing disease early.

3.2 Objective

The objective of this paper is to investigate how supervised Machine Learning (ML) algorithms can enhance healthcare by enabling more precise and early detection of diseases. In order to achieve this, we will evaluate research studies that employ multiple supervised ML models for each disease recognition task. By using a variety of algorithms in our analysis, we can obtain more comprehensive and accurate results. This approach helps to mitigate biases that may arise from evaluating a single algorithm across different research scenarios, which can lead to misleading conclusions.

3.3 Models used in proposed method

There are various methodologies that can be adapted to satisfy the objective of disease prediction. Here are some of them:

- Machine learning algorithms Machine learning algorithms can be trained on a dataset of features such as demographic information, medical history, lifestyle factors, and biomarkers to predict the likelihood of an individual having a particular disease. There are various machine learning algorithms such as logistic regression, random forests, and neural networks that can be used for disease prediction.
- Risk scores Risk scores are widely used to predict the likelihood of developing a particular disease. These scores are usually calculated based on a set of risk factors such as age, sex, family history, and lifestyle factors. For example, the Gail model is used to predict the risk of breast cancer, and the Framingham risk score is used to predict the risk of cardiovascular disease.
- Decision trees Decision trees are a type of algorithm that can be used to predict the likelihood of a particular disease based on a set of symptoms and risk factors. Decision trees are particularly useful when the data is structured and can help identify the most important factors for disease prediction.
- Bayesian networks Bayesian networks are a probabilistic graphical model that can be used to represent the relationships between different diseases and risk factors. Bayesian networks can be used for disease prediction by incorporating prior knowledge about the relationships between diseases and risk factors and then predicting the probability of an individual developing a particular disease.
- Deep learning Deep learning is a subset of machine learning that uses neural networks to extract features from the data. Deep learning algorithms can be used for disease prediction by training on large datasets of features such as medical images, electronic health records, and genomic data.
- These methodologies can be adapted to different diseases and datasets to predict the likelihood of an individual developing a particular disease. It is important to note that these methodologies require large datasets and expert knowledge to ensure accurate disease prediction.

4 Proposed method

The Multiple Disease Prediction System using Machine Learning is an advanced healthcare application that utilizes machine learning algorithms to accurately predict the likelihood of a patient having multiple diseases based on their medical history and symptoms. The system employs a comprehensive dataset of medical records and symptoms of various diseases, which are then analysed using machine learning techniques such as decision trees, support vector machines, and random forests. The system's predictions are highly accurate, and it can assist medical professionals in making more informed decisions and providing better treatment plans for patients. Ultimately, the Multiple Disease Prediction System using Machine Learning has the potential to improve healthcare outcomes and reduce healthcare costs by predicting and preventing disease early.

4.1 Architecture diagram of the proposed work

The process of disease prediction involves several steps that are designed to identify the likelihood of an individual developing a particular disease. It starts with data collection, where relevant information about the individual is collected, including their medical history, lifestyle factors, demographic information, and biomarkers. Once the data is collected, it is preprocessed to ensure that it is suitable for analysis. This may include cleaning the data, handling missing values, and normalizing the data. Feature selection is then performed to identify the most important features relevant to disease prediction. This helps to reduce the dimensionality of the data and improve the accuracy of the prediction.

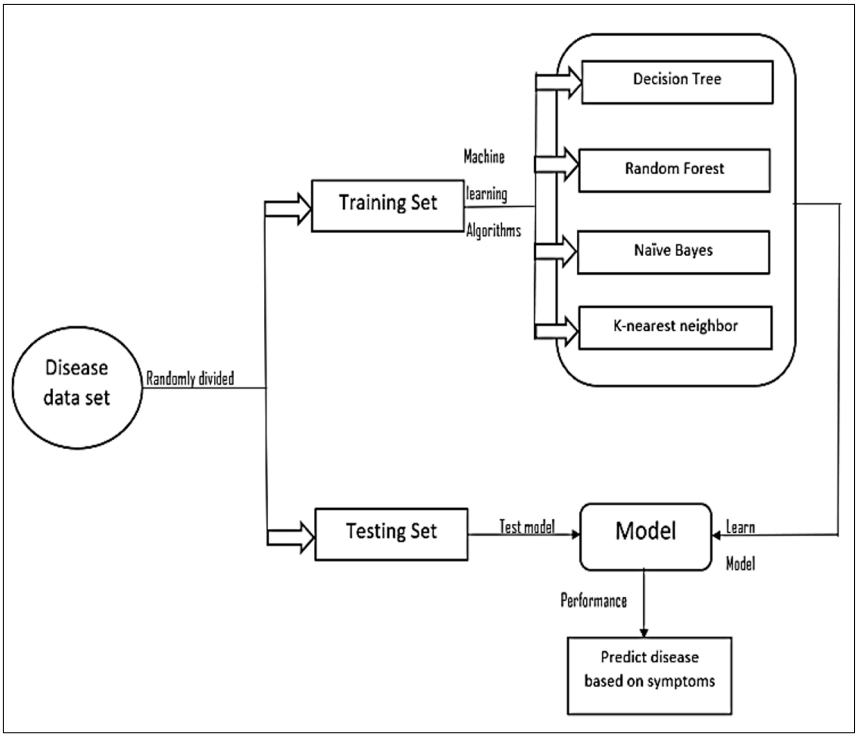

Fig. 1. Image deblurring architecture diagram.

Next, an appropriate model is selected for disease prediction. This can involve choosing a machine learning algorithm, decision tree, or Bayesian network based on the type of data and disease being predicted. Once the model is selected, it is trained on the data to identify patterns and relationships between the features and the disease being predicted. Model evaluation is then performed to test the accuracy and generalization performance of the trained model on a separate dataset. Finally, the trained model is used to predict the likelihood of an individual developing the disease. Overall, the disease prediction process requires expertise in data science, machine learning, and medical domain knowledge to ensure accurate disease prediction.

5 Results and discussions

5.1 Description about dataset

In this paper we are using multiple datasets for training our model, these multiple datasets are diabetes, heart, kidney, liver, and cancer and are downloaded from the web. These multiple datasets contain multiple attributes among those attributes only attributes that help in predicting disease are considered, and which help in increasing accuracy are considered. The datasets used in this paper are:

- Diabetes Dataset Commonly used diabetes attributes are No. Of pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age.
- Breast Cancer Dataset Attributes helped in breast cancer detection are ID, Diagnosis, Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave Points, Symmetry, Fractal Dimension, area worst, and symmetry worst.
- Heart Disease Dataset Commonly used attributes are Age, Sex, Chest Pain Type, Resting Blood Pressure, Cholesterol, Fasting Blood Sugar, Resting Electrocardiographic Results, Maximum Heart Rate Achieved, Exercise Induced Angina, ST Depression Induced by Exercise, Slope of the Peak Exercise ST Segment, Number of Major Vessels (0-3) Colored by Fluoroscopy, Thalassemia, Target Variable.
- Kidney Disease Prediction Attributes helped in kidney disease detection are Age, Blood Pressure, Specific Gravity, Albumin, Sugar, Red Blood Cells, Pus Cell, Pus Cell Clumps, Bacteria, Blood Glucose Random, Blood Urea, Serum Creatinine, Sodium, Potassium, Hemoglobin, Packed Cell Volume, White Blood Cell Count, Red Blood Cell Count, Hypertension, Diabetes Mellitus, Coronary Artery Disease, Appetite.
- Liver Dataset Commonly used attributes in liver disease prediction are Age, Total Bilirubin, Alkaline Phosphate, Aspartate Aminotranferase, Albumin, Gender, Conjugated Bilirubin, Alamin Aminotransferase, Total Protein, Albumin and Globulin Ratio.

5.2 Experimental results

The accuracy of Diabetes prediction model performed is given by ROC curve, among all the models Random Forest gave an accuracy of 92.5%.

The actual web site that shows details about multiple disease, the web page include details about the disease and symptoms related to disease. The patient now can choose his preferred disease at top right and start entering values in the text boxes, diabetes web page where user have to enter values in the text boxes. The breast cancer web page where user

have to enter values in the text boxes. The user has to enter detailed values in the text boxes, the user then clicked predict to get the results. It will redirect to result page.

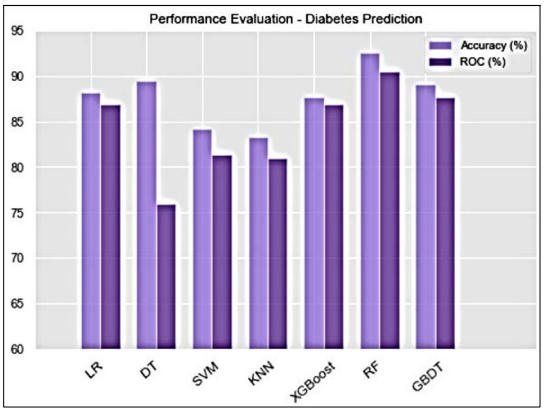

Fig. 2. Diabetes performance evaluation.

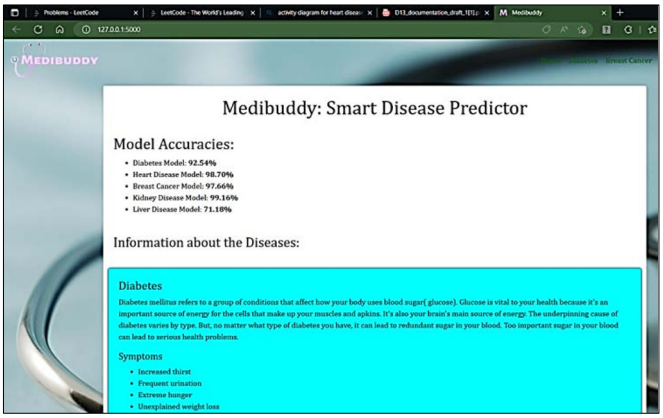

Fig. 3. Home page.

Fig. 4. Diabetes page.

Fig. 5. Inserting values.

Figure 5 represents results web page where user will get the results if the user is facing any problem then the result page will display positive, the patient may be facing disease else negative the patient is safe and also displays home button below. The figure 6 represents the result page for diabetes. In this way for every page, the result page will be displayed as shown above.

Fig. 6. Result page.

5.3 Significance of proposed method

Predicting multiple diseases using dedicated models offers several advantages:

1. Comprehensive risk assessment multiple disease prediction models consider a wide range of risk factors and indicators associated with different diseases. By combining these factors, they provide a more comprehensive evaluation of an individual's overall disease risk. This allows for a more accurate assessment compared to individual disease-specific models.
2. Holistic approach to healthcare multiple disease prediction models promote a holistic approach to healthcare by considering the interplay between different diseases. They

take into account the potential correlations, shared risk factors, and interactions among diseases, providing a more accurate representation of an individual's health status.

3. Early detection of interconnected diseases often share common risk factors and may co-occur in individuals. Multiple disease prediction models can identify such interrelationships, enabling early detection of associated diseases. This early detection facilitates timely intervention, leading to improved outcomes and potentially preventing the progression of multiple diseases.
4. Efficient use of data and resources Developing multiple disease prediction models allows for the integration and utilization of diverse datasets and resources. By leveraging shared data and risk factors, researchers can optimize the use of available resources, such as large-scale population health data or electronic health records, to develop accurate and efficient models.
5. Personalized prevention and intervention strategies Multiple disease prediction models enable the development of personalized prevention and intervention strategies. They can identify individuals at high risk for multiple diseases and provide tailored recommendations for risk reduction, lifestyle modifications, and targeted interventions. This personalized approach enhances the effectiveness of preventive measures and interventions.
6. Improved decision-making for healthcare providers Multiple disease prediction models provide healthcare providers with comprehensive information to make informed decisions. By considering the risks of multiple diseases, healthcare professionals can prioritize preventive measures, screenings, and interventions based on an individual's overall disease risk. This aids in optimizing healthcare management and resource allocation.
7. Insights into disease interactions and comorbidities Multiple disease prediction models contribute to a better understanding of disease interactions, comorbidities, and their underlying mechanisms. They help identify common pathways, shared risk factors, and potential synergistic effects between diseases. This knowledge can lead to advancements in disease prevention, treatment, and the development of targeted therapies.

In summary, multiple disease prediction models offer a holistic and integrated approach to disease risk assessment, enabling early detection, personalized interventions, and efficient resource allocation. They provide valuable insights into disease interrelationships, promoting better healthcare decision-making and improving health outcomes for individuals at risk for multiple diseases.

6 Conclusion

The primary objective of this paper is to automatically predict diseases accurately based on patient-reported symptoms by implementing Machine Learning algorithms. In this study, four Machine Learning algorithms were utilized, achieving a mean accuracy of over 95%. This signifies significant improvement and higher accuracy compared to previous works, making the system more reliable and satisfying for users.

Overall, the integration of Machine Learning in disease prediction has the potential to revolutionize healthcare by improving prediction accuracy, enabling early interventions, facilitating personalized medicine, optimizing resource allocation, and generating data-driven insights. Continued research, collaboration between healthcare professionals and data scientists, and ethical considerations are essential for harnessing the full potential of Machine Learning in disease prediction and providing enhanced healthcare solutions for

patients. Additionally, this system is user-friendly and accessible to a wide range of users without any specific threshold.

References

1. S. Khurana, A. Jain, S. Kataria, K. Bhasin, S. Arora, A. D. Gupta, Intl. Res. J. Engg. Tech **6**, 5 (2019)
2. Kamboj, Intl. J. Sci. Res **9**, 7 (2020)
3. Ware, Rakesh, Choudhary, Intl. J. Rec. Tech. Engg **8**, 5 (2020)
4. Shirsath, Patil, Intl. J. Innov. Res. Sci. Tech **7**, 6 (2018)
5. Marimuthu, Intl. J. Comp. Appl **181**, 18 (2018)
6. Battineni, Intl. J. Person. Med **10**, 21 (2020)
7. Ardabili, J. Algor. **13**, 249 (2020)
8. Shrestha, Chatterjee, LBEF Res. J. Sci. Tech. Manag **1**, 2 (2019)
9. J. Magar, Emerg. Technol. Innov. Res **7**, 6 (2020)
10. Alotaibi, Int. J. Adv. Comput. Sci. Appl **10**, 6 (2019)
11. Godse, Int. J. Adv. Res. Comput. Commun. Eng **8**, 12 (2019)
12. Anitha, Sridevi, J. Anal. Comput **13**, 2 (2019)
13. Bindhika, Int. Res. J. Eng. Technol **7**, 4 (2020)
14. Pingale, Int. Res. J. Eng. Technol **6**, 12 (2019)
15. Chauhan, Int. Res. J. Eng. Technol **7**, 1 (2020)
16. R. P. Ram Kumar, P. Sanjeeva, S. F. Lazarus, D. V. Krishna, Intl. J. Inno. Tech. Explor. Engg 8, 11S2 (2019)
17. M. Thejaswee, V. Srilakshmi, K. Anuradha, G. Karuna, *Performance Analysis of Machine Learning Algorithms for Text Classification*, in Proceedings of the Advanced Informatics for Computing Research (ICAICR 2020), A. K. Luhach, D. S. Jat, K. H. Bin Ghazali, Gao, P. Lingras, (eds), Comm. Comp. Inform. Sci. Springer, Singapore 1393 (2021)
18. B. Sankara Babu, S. Nalajala, K. Sarada, V. Muniraju Naidu, N. Yamsani, K. Saikumar, Machine Learning based online Handwritten Telugu Letters Recognition for Different Domains, in Proceedings of A Fusion of Artificial Intelligence and Internet of Things for Emerging Cyber Systems, P. Kumar, A. J. Obaid, K. Cengiz, A. Khanna, V. E. Balas (eds), Intelligent Systems Reference Library, vol 210. Springer, (2022)
19. R. P. Ram Kumar, R. Tabassum, Intl. J. Creat. Res. Thoug **6**, 1 (2018)
20. A. Sankaridevi, R. P. Ram Kumar, R. Jayakumar, Intl. J. Recen. Tech. Engg **7**, 5C (2019)
21. R. P. Ram Kumar, S. Polepaka, Performance Comparison of Random Forest Classifier and Convolution Neural Network in Predicting Heart Diseases, in Proceedings of the Third International Conference on Computational Intelligence and Informatics, (eds) K. Raju, A. Govardhan, B. Rani, R. Sridevi, M. Murty, Advances in Intelligent Systems and Computing, 1090. Springer, Singapore (2020)