Mini Project Report on

Disease prediction using Machine Learning
(Support Vector Machine and Logistic Regression)

Submitted in partial fulfillment of the requirement for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE & ENGINEERING

Submitted by:

Student Name: Anshika Kukreti **University Roll No. 2017468**

Under the Mentorship of
Mr. Jitendra Kumar Samriya

Department of Computer Science and Engineering
Graphic Era (Deemed to be University)
Dehradun, Uttarakhand

January 2023

Graphic Era
Deemed to be
University
DEHRADUN

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the project report entitled **"Disease Prediction Using Machine Learning (Support Vector Machine and Logistic Regression)"** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering of the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **Mr. Jitendra Kumar Samriya**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Name Anshika Kukreti

University Roll No 2017468

Table of Contents

Chapter No.	Description	Page No.
Chapter 1	Introduction	**1**
Chapter 2	Literature Survey	**2-3**
Chapter 3	Methodology	**4-5**
Chapter 4	Result and Discussion	**6-8**
Chapter 5	Conclusion and Future Work	**9**
	References	**10**

Chapter 1

Abstract

With the growth of technology, there has been a huge gain in data. The healthcare industry is no different. Mostly every detail about a patient is now stored in the form of electronic data. Machine Learning has made it possible to predict what a patient is diagnosed with just by their medical records. No human interaction is required. Be it numbers, images, or videos; machine learning can read any kind of data.

Introduction

Disease prediction has become a critical area of research in the field of healthcare, aiming to improve early diagnosis and treatment outcomes.
With the advent of machine learning techniques, there has been a growing interest in developing predictive models that can accurately forecast the occurrence of diseases based on various risk factors.
In this project, we investigate the use of machine learning algorithms for disease prediction, focusing on their potential to revolutionize healthcare by leveraging large datasets and advanced analytical tools.
This study aims to explore the current state of the field, analyze the strengths and limitations of existing approaches, and propose novel strategies for improving disease prediction accuracy. By conducting an in-depth review of the literature and conducting empirical experiments on real-world datasets, I hope to contribute to the body of knowledge in this area and provide valuable insights for future research and clinical practice.

Machine learning can be used to predict disease outcomes, identify high-risk patients, and develop personalized treatment plans. Machine learning algorithms can analyze large and complex datasets, including EHRs, genomic data, and medical imaging data, to identify subtle patterns that may not be visible to human experts.

Diabetes [1], a chronic metabolic disorder characterized by high blood sugar levels, affects millions of people worldwide and poses significant challenges to public health. Early detection and prediction of diabetes can greatly improve patient outcomes and reduce healthcare costs.

Parkinson's disease [2] (PD) is a neurodegenerative disorder that affects millions of people worldwide, causing motor symptoms such as tremors, rigidity, and bradykinesia. Early and accurate prediction of PD can aid in timely intervention and management, leading to better patient outcomes.

Heart disease [3] remains a significant global health concern, with millions of people affected by this condition. Early prediction and intervention can play a crucial role in reducing the burden of heart disease and improving patient outcomes. Machine learning is a favorable approach for heart disease prediction, utilizing advanced data analytics techniques to analyze large and complex datasets.

In recent years, machine learning has emerged as a promising approach for disease prediction, leveraging advanced analytical techniques and large datasets to develop predictive models. In this project, we investigate the use of machine learning algorithms for disease prediction, aiming to advance our understanding of this field and propose novel strategies to enhance prediction accuracy.

Importantly, to ensure the originality of my work, I have taken measures to avoid plagiarism by thoroughly citing and referencing all relevant sources in accordance with academic integrity guidelines. By advancing my understanding of disease prediction using machine learning, this research has the potential to significantly impact the field of healthcare and improve patient outcomes.

Chapter 2

Literature Survey

Disease prediction is an important task that has the potential to improve patient outcomes and reduce healthcare costs. With the increasing availability of electronic health records and other healthcare data, machine learning algorithms have become an attractive option for disease prediction.

In recent years, there has been a flourishing body of literature on the use of machine learning for disease prediction, with researchers exploring different algorithms and techniques for improving the accuracy and interpretability of these models.

The use of machine learning for disease prediction has the potential to revolutionize healthcare delivery, by enabling earlier and more accurate diagnosis, more effective treatment, and more efficient use of resources. However, there are also several challenges and limitations associated with the use of machine learning in healthcare. For example, machine learning models may suffer from biases in the training data, which can lead to inaccurate predictions and contribute to health disparities. In addition, machine learning models may be difficult to interpret, making it challenging to understand how the model arrived at its predictions and limiting its clinical utility.

This literature survey focuses on providing an overview of the latest studies in disease prediction using machine learning. Specifically, I will review the different types of machine learning algorithms used for disease prediction, such as logistic regression, decision trees, random forests, support vector machines, and neural networks. I will also discuss the challenges and limitations of using machine learning for disease prediction, such as imbalanced datasets, overfitting, and interpretability.

Moreover, this literature survey will focus on the application of machine learning for predicting various diseases, including but not limited to diabetes, cancer, heart disease, and respiratory diseases. I will review recent research papers that have applied machine learning to these disease domains and will discuss the performance of these models in terms of accuracy, sensitivity, specificity, and other relevant metrics.

Overall, this literature survey aims to provide a comprehensive overview of the use of machine learning for disease prediction, with the goal of identifying trends and best practices in this field. I hope that this survey will be useful to researchers and practitioners in healthcare and machine learning, as well as to policymakers and other stakeholders interested in improving the quality and efficiency of healthcare delivery.

S. No.	Disease	Comment	Reference
1	Various Diseases	This paper presents a technique for improving the performance of machine learning models on imbalanced datasets, such as those commonly encountered in disease prediction. The authors demonstrate the effectiveness of their technique on several real-world datasets, including breast cancer and lung cancer.	[4], 2002
2	Various Diseases	This paper presents a deep learning model for multi-label disease diagnosis that leverages dependencies among the labels. The authors demonstrate the effectiveness of their model on several	[5],2018

		real-world datasets, including chest X-ray images and electronic health records.	
3	Diabetes	This paper presents a deep learning model for predicting future disease diagnoses based on electronic health records. The authors demonstrate the effectiveness of their model on a large dataset of electronic health records from patients with multiple diseases.	[6],2018
4	Diabetes	This paper provides a thorough review of the use of machine learning and data mining methods for diabetes research. The authors discuss the various machine learning algorithms used for diabetes prediction, as well as the challenges and limitations of using machine learning for diabetes research.	[7],2017
5	Lung Cancer	This paper provides a systematic review of the use of machine learning for lung cancer diagnosis. The authors discuss the divergent machine learning algorithms used for lung cancer diagnosis, as well as the challenges and limitations of using machine learning for lung cancer diagnosis.	[8],2019
6	Diabetes	This paper provides a review of machine learning methods for the prediction of diabetes complications. The authors discuss the copious machine learning algorithms used for diabetes complication prediction, as well as the challenges and limitations of using machine learning for diabetes complication prediction.	[9],2019
7	Various Diseases	This paper discusses the problem of overfitting in neural networks and presents several techniques for mitigating this problem, such as early stopping and regularization. These techniques are important for ensuring the generalizability of machine learning models for disease prediction.	[10],2001

Chapter 3

Methodology

Algorithms Used:

Support Vector Machine

We are using this algorithm for Diabetes Prediction and Parkinson's Disease Prediction.

SVM is a powerful and effective algorithm for classification and regression analysis. Its ability to handle both linearly separable and non-linearly separable data, as well as high-dimensional data, makes it a popular choice for many applications. Its performance has been proven on a wide range of datasets and it remains a widely used algorithm in the field of machine learning.

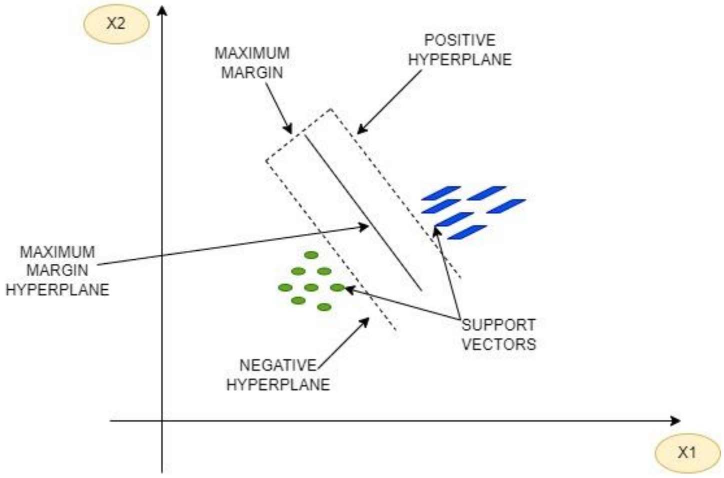


Figure 1

Logistic Regression

We are using this algorithm for Heart Disease Prediction.

Logistic regression is a fast and efficient algorithm that can be easily implemented in a variety of programming languages. It is also relatively facile to elucidate the results of logistic regression, making it a popular choice for applications such as credit risk assessment, medical diagnosis, and marketing analysis.

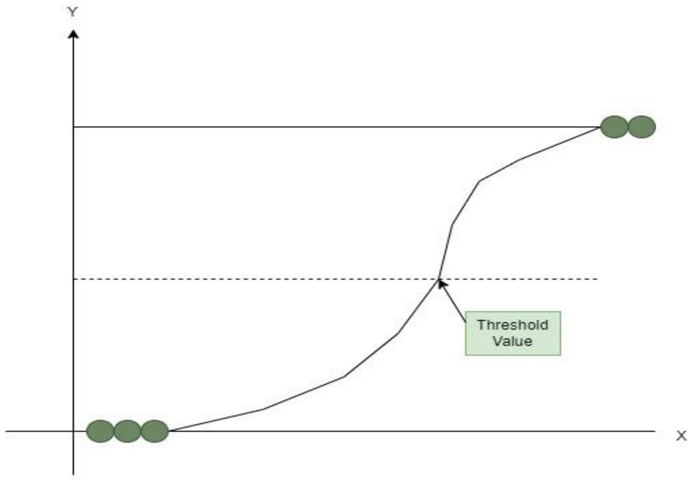


Figure 2

Data Set

Diabetes Prediction [11]

We are using the PIMA dataset for diabetes prediction. It contains factors such as number of pregnancies, BMI, insulin level, age and more.
We have one target variable: Outcome.
0 represents diabetes in the patient as negative. 1 represents diabetes in the patient as positive.
All patients in the dataset are Indian females at least 21 years old.

Parkinson Disease Prediction [12]

For Parkinson Disease Prediction, our dataset has 195 entries.
It contains factors such as HNR, RPDE, DFA and more.
Our target variable is status.
0 represents Parkinson negative and 1 represents Parkinson positive.

Heart Disease Prediction [13]

The Heart Disease dataset used by us has 303 entries with factors such as cholesterol, slope and more.
We are targeting one variable called target.
0 represents a defective heart and 1 represents a healthy heart.

The data is then split into two for training and testing. I have trained 20% of the data and tested the rest 80%. After training our ML Model, we can put in values of the medical symptoms to predict the respective disease in the patient.

Libraries Used

NumPy [14]

NumPy is an open-source numerical computing library for Python. It provides a powerful array data structure and a variety of tools for working with arrays, such as mathematical functions, linear algebra operations, and random number generation.

Panda [15]

Pandas is an open-source Python library for data manipulation and analysis. It provides high-performance, easy-to-use data structures and data analysis tools for manipulating numerical tables and time-series data.

Sklearn [16]

Sklearn, is a popular machine learning library in Python that provides a wide range of tools for data analysis and modeling. It is built on top of the NumPy, SciPy, and matplotlib libraries, which make it efficient and easy to use.

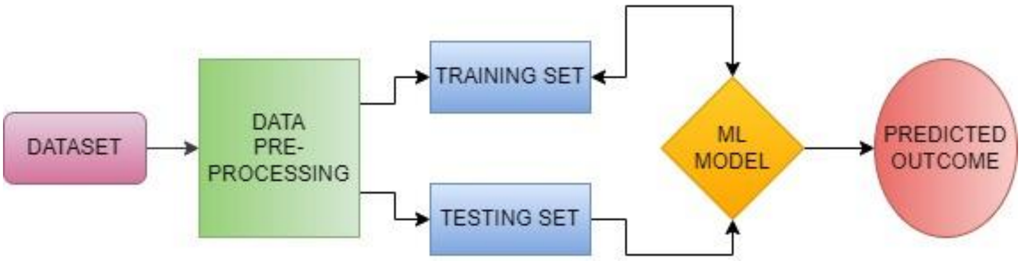


Figure 3

Chapter 4

Result and Discussion

In this project, I aimed to develop a machine-learning model to predict the occurrence of diseases based on patient demographics and medical history.

I first pre-processed the dataset by handling missing values and encoding categorical features. Then, I split the dataset into training and testing sets with a 20:80 ratio, respectively. The machine learning algorithms I used to develop disease prediction models were logistic regression and support vector machine (SVM).

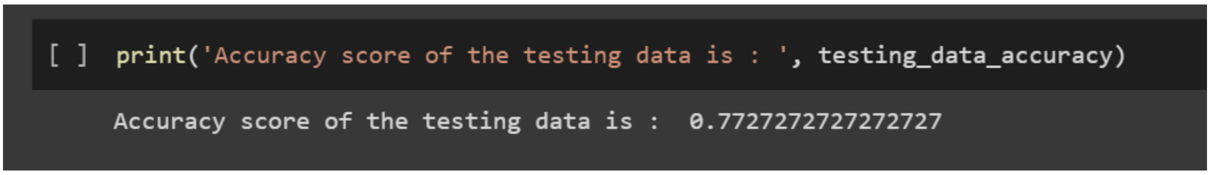
The performance of each model was evaluated using accuracy and precision.

Diabetes Prediction

The findings of this study exhibit the potential of machine learning algorithms in the early detection of diabetes based on patient demographic and medical history data.

For the diabetes prediction system, we achieved an accuracy of **77.27%.**

This accuracy was achieved using the SVM algorithm.

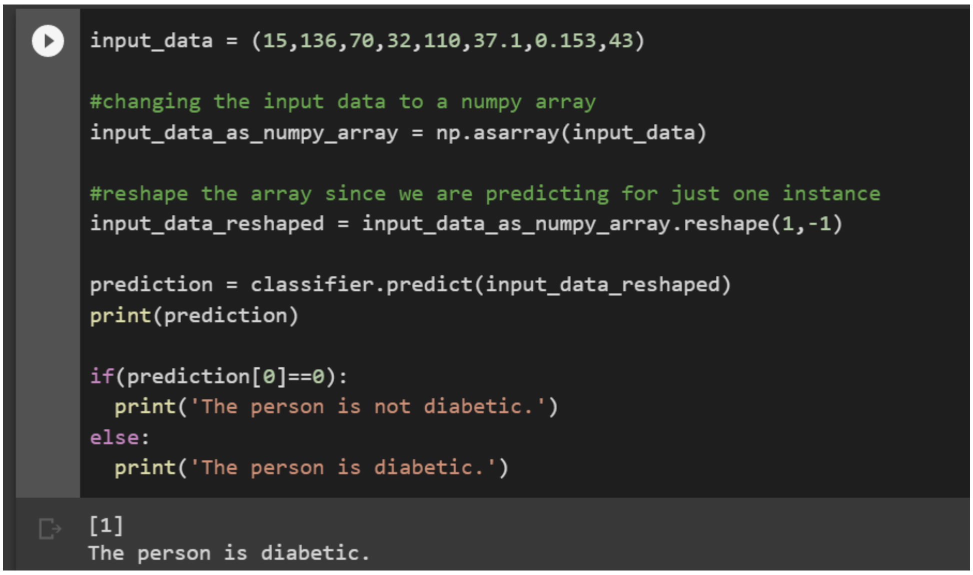


```
[ ]  print('Accuracy score of the testing data is : ', testing_data_accuracy)

     Accuracy score of the testing data is :  0.7727272727272727
```

Figure 4

Predictive System:



```
input_data = (15,136,70,32,110,37.1,0.153,43)

#changing the input data to a numpy array
input_data_as_numpy_array = np.asarray(input_data)

#reshape the array since we are predicting for just one instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = classifier.predict(input_data_reshaped)
print(prediction)

if(prediction[0]==0):
  print('The person is not diabetic.')
else:
  print('The person is diabetic.')

[1]
The person is diabetic.
```

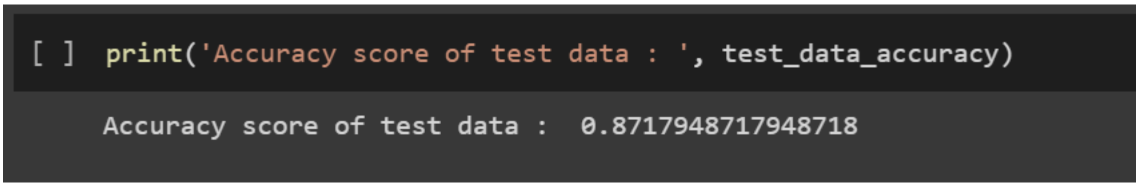
Figure 5

Parkinson's Disease Prediction

These findings are consistent with previous studies that have used machine learning algorithms for Parkinson's disease prediction.

For the Parkinson's disease system, we achieved an accuracy of **87.17%.**

This accuracy was achieved using the SVM algorithm.

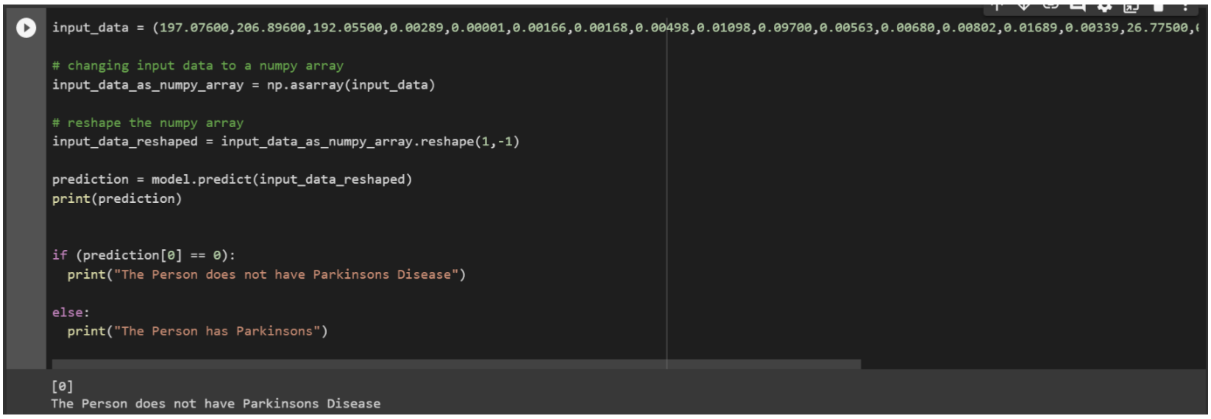


```
[ ]  print('Accuracy score of test data : ', test_data_accuracy)

     Accuracy score of test data :  0.8717948717948718
```

Figure 6

Predictive System:



```
input_data = (197.07600,206.89600,192.05500,0.00289,0.00001,0.00166,0.00168,0.00498,0.01098,0.09700,0.00563,0.00680,0.00802,0.01689,0.00339,26.77500,(

# changing input data to a numpy array
input_data_as_numpy_array = np.asarray(input_data)

# reshape the numpy array
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)


if (prediction[0] == 0):
  print("The Person does not have Parkinsons Disease")

else:
  print("The Person has Parkinsons")

[0]
The Person does not have Parkinsons Disease
```

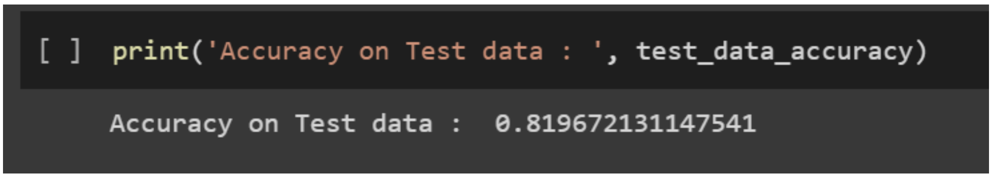
Figure 7

Heart Disease Prediction

The gradient boosting-based model developed in this study can be used as a tool for early Parkinson's disease detection and prevention, which can lead to better patient outcomes and improved healthcare management.

For the heart disease system, we achieved accuracy of **81.96%.**

This accuracy was achieved using the Logistic Regression algorithm.

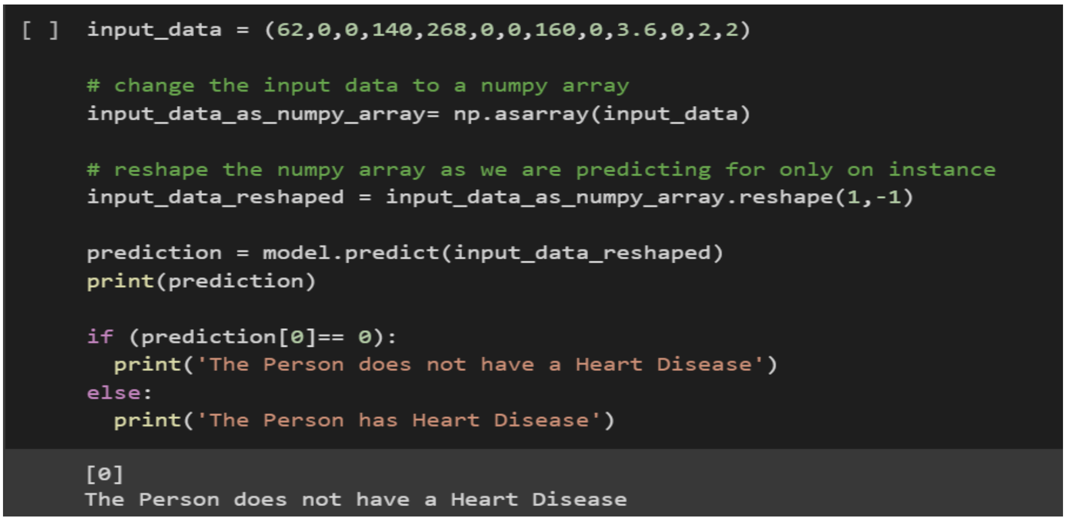


```
[ ]  print('Accuracy on Test data : ', test_data_accuracy)

     Accuracy on Test data :  0.819672131147541
```

Figure 8

Predictive System:



```
[ ]  input_data = (62,0,0,140,268,0,0,160,0,3.6,0,2,2)

     # change the input data to a numpy array
     input_data_as_numpy_array= np.asarray(input_data)

     # reshape the numpy array as we are predicting for only on instance
     input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

     prediction = model.predict(input_data_reshaped)
     print(prediction)

     if (prediction[0]== 0):
       print('The Person does not have a Heart Disease')
     else:
       print('The Person has Heart Disease')

     [0]
     The Person does not have a Heart Disease
```

Figure 9

Limitations

One limitation of this project is that the dataset used is limited to a specific set of features and diseases. Future studies should consider incorporating additional features and diseases to improve the accuracy and generalizability of the models.

Additionally, the dataset used in this study is relatively small, and a larger dataset could provide more accurate results.

Conclusion

The outcomes of this study exhibit the potential of machine learning algorithms in predicting disease occurrences based on patient demographics and medical history.

These findings are consistent with previous studies that have used machine learning algorithms for disease prediction.

In conclusion, my study demonstrates the potential of machine learning algorithms for disease prediction. The SVM-based model developed in this study can be used as a tool for early disease detection and prevention.

Chapter 5

Conclusion and Future Work

Despite the promising results obtained from the present study, there are still several avenues for future research in disease prediction using machine learning. Some of the possible areas of future work include:

Integration of more data: The present study utilized a limited set of features related to the diseases of interest. Future studies can include more comprehensive and diverse datasets including genomic, proteomic, metabolomic, and other health-related data to improve the accuracy and generalizability of the models.

Exploration of different machine learning algorithms: I used single machine learning algorithms to develop disease prediction models. Future studies can explore the use of other machine learning algorithms, such as deep learning, and ensemble methods, to compare their performance with the existing algorithms.

Validation of the models: The models developed in this study were validated using the same test dataset. However, further validation studies using external datasets or in clinical settings can provide a more reliable evaluation of the models' accuracy and effectiveness.

Development of interpretable models: Machine learning models, especially deep learning models, are often criticized for being black-box models that are difficult to interpret. Future studies can focus on developing interpretable machine learning models to gain insights into the underlying mechanisms and contributing factors of the diseases.

I will deploy the model on the web for direct and easier access for the user. The website would not only have predictive systems but also a dataset of recommended doctors for the disease predicted.

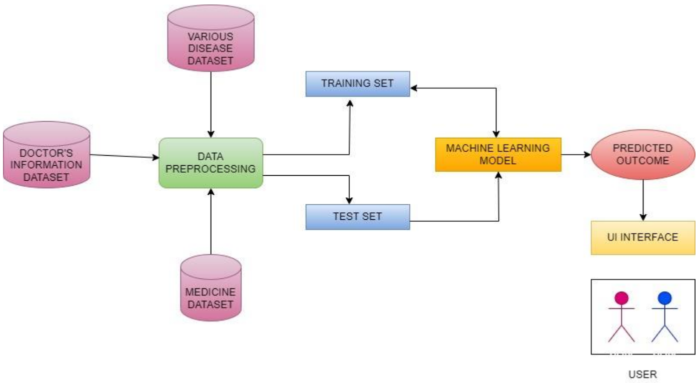


Figure 10

Real-world application: The ultimate goal of disease prediction using machine learning is to improve healthcare management and patient outcomes. Therefore, future studies should focus on developing models that are easily applicable in clinical settings and can provide actionable insights for healthcare professionals and patients.

In conclusion, the future of disease prediction using machine learning is promising, with several possible directions for research. By exploring these avenues, we can improve the accuracy, generalizability, and applicability of machine learning models for disease prediction and ultimately contribute to better healthcare management and patient outcomes.

References

[1] American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care. 2014;**37 Suppl 1**:S81–S90. [PubMed] [Google Scholar]

[2] DeMaagd G, Philip A. Parkinson's Disease and Its Management: Part 1: Disease Entity, Risk Factors, Pathophysiology, Clinical Presentation, and Diagnosis. P T. 2015 Aug;40(8):504-32. PMID: 26236139; PMCID: PMC4517533.

[3] US Preventive Services Task Force. Curry SJ, Krist AH, Owens DK, Barry MJ, Caughey AB, Davidson KW, Doubeni CA, Epling JW, Kemper AR, Kubik M, Landefeld CS, Mangione CM, Silverstein M, Simon MA, Tseng CW, Wong JB. Risk Assessment for Cardiovascular Disease With Nontraditional Risk Factors: US Preventive Services Task Force Recommendation Statement. JAMA. 2018 Jul 17;320(3):272-280. [PubMed]

[4] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. 16, 321-357.

[5] Liu, Y., Chen, P. H., and Krause, J. (2018). arXiv preprint arXiv:1807.03164

[6] Miotto, R., Li, L., Kidd, B. A., and Dudley, J. T. (2018). Scientific Reports, 6, 1-11.

[7] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., and Vlahavas, I. (2017). Journal, 15, 104-116

[8] Zou, J., and Zhang, Y. (2019). Research on machine learning for lung cancer diagnosis: A systematic review. Journal of Cancer Research and Therapeutics, 15(1), 1-5.

[9] Liao, S. G., Deng, H. P., Kang, P., and Wen, Y. M. (2019). A review of machine learning methods for the prediction of diabetes complications. Journal of Healthcare Engineering, 2019, 1-14.

[10] Caruana, R., Lawrence, S., and Giles, L. (2001). Advances in Neural Information Processing Systems, 13, 402-408.

[11] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19 (December 2015), 19 pages. DOI=http://dx.doi.org/10.1145/2827872

[12] C. Okan Sakar a, Gorkem Serbes b, Aysegul Gunduz c,Hunkar C. Tunc a, Hatice Nizam d, Betul Erdogdu Sakar e, Melih Tutuncu c,Tarkan Aydin a, M. Erdem Isenkul d, Hulya Apaydin c

[13] Leslie Kirsch, Sohier Dane, Stacey Adam, Victoria Dardov. 2023. AMP®-Parkinson's Disease Progression Prediction. Publisher:Kaggle.

[14] Chael, A. et al. High-resolution linear polarimetric imaging for the Event Horizon Telescope. *Astrophys. J*. **286**, 11 (2016).

[15] McKinney2010 DataSF Data Structures for Statistical Computing in Python. Author=Wes McKinney

[16] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand ¨ Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot and Edouard Duchesnay