IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

CHRONIC DISEASES PREDICTION AND DETECTION USING MACHINE LEARNING

[1]Abhinandan Katoch, [2]Nitin Choudhary, [3]Oshin Sharma

[1]Student, [2]Student, [3]Assistant Professor
[1]Department of Computer Science and Engineering,
[1]SRMIST Delhi – NCR Campus, Ghaziabad, India

Abstract: A Machine Learning based approach for the prediction and detection of Chronic Diseases such as “Type-2-Diabetes, Parkinson’s Disease, Heart Disease and Classification of Type of Breast Cancer”. The developed machine learning models can predict the likelihood of a person developing such Chronic Ailment based on demographic and health-related data. The models work on two main Machine Learning algorithms: Logistic Regression and Decision Tree Classifier which are trained on a fine and large number of datasets. All of the backend functions are then reflected on frontend using a Web Application with simple user interface.

***Index Terms* - Machine Learning, Chronic Diseases, Logistic Regression, Decision Tree.**

I. INTRODUCTION

Chronic Diseases are major concern these days as they have serious consequences on a person’s health and well-being. These conditions are typically long-lasting and can worse over time, leading to disability, reduced quality of life and even fatality. As the number of chronic diseases keeps rising, “India is going through a rapid shift in terms of health. 53% of fatalities and 44% of DALYs (Disability Adjusted Life Years) lost in 2005 are thought to have been caused by these illnesses”. Previous projections from the “Global Burden of Disease Study” indicated that “The number of deaths attributable to chronic diseases would increase from 3.78 million in 1990 (which accounted for 40.4% of all deaths) to 7.63 million in 2020 (or 66.7% of all deaths)”. [1]

The primary causes of death in this rural area of India are chronic diseases. It is probable that this pattern of mortality is not exclusive to these specific villages, and it offers novel understanding regarding the swift advancement of epidemiological transition in rural India. [2]

Early detection and prevention of Chronic Diseases can lead to better treatment outcomes and improved quality of life for patients. Machine Learning has shown promising results in such situations leveraging the power of computational algorithms. Machine learning algorithms can detect trends and forecast which therapies are most likely to be successful for certain patients by analysing vast volumes of patient data. Therefore, the web application that we have created has the potential to aid healthcare professionals in detecting Chronic Diseases even beyond the confines of hospitals and clinics. By inputting a medical diagnosis report into the application, which includes attributes specific to the targeted disease, the web application can utilize dedicated models in the backend to process the results.

This means that healthcare professionals can utilize the web application to diagnose Chronic Diseases in patients outside of traditional medical settings, providing more comprehensive and accessible healthcare. The use of dedicated models in the backend of the web application ensures that the results are accurate and reliable.

II. LITERATURE SURVEY

The study of Chronic Diseases prevention using Machine Learning is a growing topic now a days as they are responsible for the majority of deaths world-wide. These diseases are often preventable, and research has shown that early detection of such diseases increases the likelihood of a person’s surviving. The literature review we did give a critical evaluation and analysis of existing published research articles on this topic. The literary survey provided a comprehensive overview of the current state of knowledge and understanding to identify gaps, inconsistencies, and contradictions in the existing literature. Table 1 contains some recently published Articles and Research Papers on the use of Machine Learning in field of Medical Science.

Table 1. Tabular Summary of Literature Survey

Author (Year)	Title of Paper	Methodology	Dataset Used	Limitations
"Rayan Alanazi" (2022)	"Identification and Prediction of Chronic Diseases Using Machine Learning Approach" [3]	"CNN, KNN"	"Immune Epitope Database (IEDB), Nationwide Inpatient Sample (NIS), International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM)"	Narrow selection and application of machine learning models. Need large and very high-quality data. Lack of optimal datasets.
"Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang and Lin Wang" (2017)	"Disease Prediction by Machine Learning over Big Data from Healthcare Communities" [4]	"CNN-UDPR, CNN-MDRP"	Real-life Hospital Data which include EHR, Medical Image and Gene Data. Source is classified due to privacy	CNN Model/ Algorithm requires a lot of training data to be effective and here training data is always a challenge.
"Haohui Lu, Shahadat Uddin, Farshid Hajati, Mohammad Ali Moni, Khushi" (2022)	"A Patient Network-Based Machine Learning Model for Disease Prediction - The Case of Type 2 Diabetes Mellitus" [5]	"KNN, SVM, NB, DT, RF, XGBoost, ANN"	CBHS Health Funds (Australian Health Insurance Company)	The issue of strong and effective dataset still arises as here also, real-world healthcare data is used which is hard to process. Multiple information of patients were found missing.
"Poonam Sinha and Parul Sinha" (2015)	"Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM" [6]	"SVM, KNN."	Data is collected form several medical diagnostics laboratories particularly focusing on the patients result of KFT (Kidney Function Test)	Proposed solution is solely for chronic kidney diseases. Dataset should be optimal and model selection should be narrowed to one strong algorithm with maximum accuracy result.
"Gopi Battineni, Getu Gamo Sagaro, Nalini and Francesco Amenta" (2020)	"Applications of Machine Learning Predictive Models" [7]	"NB, RF, KNN, SVM, NN, ANN, DT"	"PubMed (Medline), Cumulative Index to Nursing and Allied Health Literature (CINAHL)"	There are no complete AI systems that can be fully trusted. Cost effective but still can't replace the quantified computed tomography and imaging.

“R. L. Priya and S. Vinila Jinny” (2021)	“Elderly Healthcare System for Chronic Ailments using Machine Learning Techniques” [8]	CNN	Clinical Notes from Hospitals and Health Insurance Companies, X-Ray and Radiology Reports	CNN Model/ Algorithm requires a lot of training data to be effective and here training data is always a challenge. Dataset should be optimal and efficient.

III. PROPOSED METHODOLOGY

The research we conducted target four different types of Chronic Diseases: “Type-2-Diabetes, Parkinson’s Disease, Heart Disease and Classification of Type of Breast Cancer [Malignant or Benign]”. The best fitted Machine Learning algorithms our research narrowed down to are Logistic Regression and Decision Tree Classifiers. The Type-2-Diabetes and Parkinson’s Disease Detection works under Decision Tree Classifiers algorithm and on the other hand Heart Disease Detection and Classification of Type of Breast Cancer works under Logistic Regression algorithm.

- Logistic Regression is a powerful statistical analysis algorithm used to model the probability of a certain event occurring. The main goal of logistic regression is to build a model that can predict a binary outcome (1 or 0) based on a set of independent variables. It works by estimating the probability of the dependent variable (the outcome) using a “Sigmoid Function”, which maps any real-valued number to a value between 0 and 1. Logistic Regression Equation:

$$\boxed{\log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n}$$

- Decision Tree Classifier is also a powerful Machine Learning algorithm which is used for both classification and regression problems. It makes a tree-like model of decisions and their possible outcomes, which helps in decision making. The tree is made by recursively dividing the data into sub-sets, with each partition being based on the value of a single feature. It starts at the root node and traverses the tree by following the decision rules at each node, until a leaf node is reached. Decision tree classifiers have advantages such as “interpretability, ability to handle both categorical and continuous features, and scalability to large datasets”.

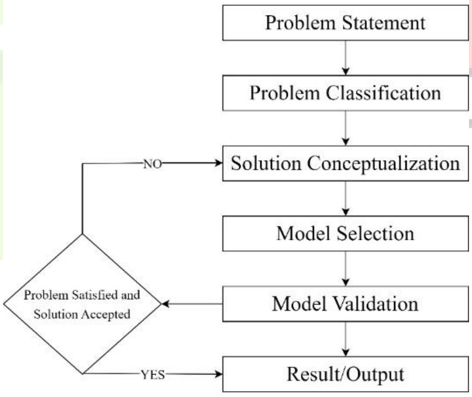

Figure 1. Proposed Methodology and System Architecture

The research conducted demands a diverse range of data, At the heart of our data sets lies the medical diagnostic reports of patients suffering from above mentioned Chronic Diseases. This data provides the foundation upon which the Machine Learning models are build and trained. Table 2 below shows the data sets used along with their sources respectively.

Table 2. Dataset and Source

Chronic Disease	Dataset Source
Type-2-Diabetes	PIMA Diabetes Dataset (NIDDK)
Heart Disease	UCI Machine Learning Archive
Parkinson’s Disease	University of Oxford & NCVS
Breast Cancer Classification	UCI Machine Learning Archive

IV. EXSITING PROBLEM AND PROPOSED SOLUTIONS

The algorithms require a significant amount of data to be accurate and can be difficult to interpret and explain. This can be a problem for medical professionals who may not have a background in machine learning and may struggle to understand the results produced by these models. The solution is to use more interpret-able models to help medical professionals.

Ethical considerations that must be taken into account when using machine learning for medical purposes. Patient privacy and safety are of utmost importance, and it can be difficult to ensure that sensitive medical data is kept secure. Use of data sets that are standardized or one that does not contain any attribute that reveals once private information such as: Ethnicity, Race, Name or Address.

V. SYSTEM ARCHITECTURE

The ultimate objective of this research is to develop and deploy fully functioning Machine Learning models as a user-friendly Web-Application for healthcare professionals. By doing so, we aim to provide a simple and accessible tool that can assist healthcare professionals in predicting and detecting chronic diseases accurately and efficiently. Developing a web app provides us with the opportunity to continuously improve and update our models to ensure their accuracy and reliability.

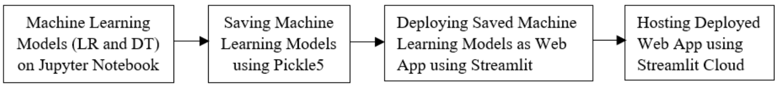
Machine Learning Models (LR and DT) on Jupyter Notebook
Saving Machine Learning Models using Pickle5
Deploying Saved Machine Learning Models as Web App using Streamlit
Hosting Deployed Web App using Streamlit Cloud

Figure 2. Web App Deployment

Thorough steps involved in the functioning of our system are: User inputs the medical diagnostic data into a simple web page. In the backend, the data is then fed to the Machine Learning model assigned with that particular disease. The model will validate the solution (which is either a positive or negative response). The result is then displayed to the front end along with an advisory notice on not fully rely on Machine Learning generated results and seeking professional medical help is always advised if symptoms get any worse.

VI. SYSTEM SPECIFICATIONS

Table 3. System Specification

Description	Type/Version
Processor (CPU)	Intel(R) Core (TM) i7-10750H
GPU	Intel(R) UHD Graphics
Disk	512 GB SSD
Memory	16 GB DDR4
External Components	Standard Keyboard and Mouse
Operating System	Windows(R) 11
Python	3.11.1
Streamlit Community Cloud	2023

VII. RESULT AND ANALYSIS

The main goal of this research paper was to sketch how Machine Learning can be show effective results in prediction and detection of Chronic Diseases as early as one experience worrying symptoms. The construction of these Machine Learning models based on Logistic Regression and Decision Tree Classifier Algorithm can assist the healthcare workers to diagnose and detect the likelihood of a person developing Chronic Diseases as we have also integrated those models into a web application.

Table 4. Result for Type – II Diabetes Prediction Model (Algorithm: Decision Tree)

Attributes	Case 1	Case 2
No. of Pregnancies	1	1
Glucose (mg/dl)	95	189
Blood Pressure (mmHg)	66	60
Skin Thickness (mm)	13	23
Insulin (mu/ml)	38	846
BMI	19.6	30.1
Diabetes Pedigree Function	0.334	0.398
Age	25	59
Output Obtained	**[0]**	**[1]**
Remarks	**Non-Diabetic**	**Diabetic**

Table 5. Result for Heart Disease Prediction Model (Algorithm: Logistic Regression)

Attributes	**Case 1**	**Case 2**
Age	43	54
Gender (1=Male, 0=Female)	0	1
Chest Pain Value	0	1
Blood Pressure (mmHg)	132	108
Cholesterol (mg/dl)	341	309
Glucose (>120 mg/dl)	1	0
Resting ECG Result	0	1
Maximum Heart Rate	136	156
Exercise Induced Angina	1	0
ST Depression Value	3	0
Peak ST Segment Value	1	2
Major Vessels (Fluoroscopy)	0	0
Thalassemia Value	3	3
Output Obtained	**[0]**	**[1]**
Remarks	**Health Heart**	**Cardiovascular Complications**

Table 6. Result for Breast Cancer Type Classification Model (Algorithm: Logistic Regression)

Attributes	**Case 1**	**Case 2**
Mean Radius	14.25	13.34
Mean Texture	21.72	15.86
Mean Perimeter	93.63	86.49
Mean Area	633	520
Mean Smoothness	0.09823	0.1078
Mean Compactness	0.1098	0.1535
Mean Concavity	0.1319	0.1169
Mean Concave Points	0.05598	0.06987
Mean Symmetry	0.1885	0.1942
Mean Fractal Dimensions	0.06125	0.06902
Radius Error	0.286	0.286
Texture Error	1.019	1.016
Perimeter Error	2.657	1.535
Area Error	24.91	12.96
Smoothness Error	0.005878	0.006794
Compactness Error	0.02995	0.03575
Concavity Error	0.04815	0.0398
Concave Points Error	0.01161	0.01383
Symmetry Error	0.02028	0.02134
Fractal Dimensions Error	0.004022	0.004603
Worst Radius	15.89	15.53
Worst Texture	30.36	23.19
Worst Perimeter	116.2	96.66
Worst Area	799.6	614.9
Worst Smoothness	0.1446	0.1536
Worst Compactness	0.4238	0.4791
Worst Concavity	0.5186	0.4858
Worst Concave Points	0.1447	0.1708
Worst Symmetry	0.3591	0.3527
Output Obtained	**[0]**	**[1]**
Remarks	**Malignant Growth of Cancer**	**Benign Growth of Cancer**

Table 7. Result for Parkinson's Disease Detection Model (Algorithm: Decision Tree)

Attributes	**Case 1**	**Case 2**
MDVP-Fo (Hz)	241.40400	139.22400
MDVP-Fhi (Hz)	248.83400	586.56700
MDVP-Flo (Hz)	232.48300	66.15700
MDVP-Jitter (%)	0.00281	0.03011
MDVP-Jitter (Abs)	0.00001	0.00022
MDVP-RAP	0.00157	0.01854
MDVP-PPQ	0.00173	0.01628
MDVP-DDP	0.00470	0.05563
MDVP-Shimmer	0.01760	0.09419
MDVP-Shimmer (dB)	0.15400	0.93000
Shimmer-APQ3	0.01006	0.05551
Shimmer-APQ5	0.01038	0.05005
MDVP-APQ	0.01251	0.06023
Shimmer-DDA	0.03017	0.16654
NHR	0.00675	0.25930
HNR	23.14500	10.48900
RPDE	0.457702	0.596362
DFA	0.634267	0.641418
Spread1	-6.793547	-3.269487
Spread2	0.158266	0.270641
D2	2.256699	2.690917
PPE	0.117399	0.444774
Output Obtained	**[0]**	**[1]**
Remarks	**Parkinson's Negative**	**Parkinson's Positive**

The prime findings we concluded at the end of this research was that use of Decision Tree Classifier was a very effective option while handling large dataset which contains mean values and data standardization is required. On the other hand, Logistic Regression also showed very promising results and a well acceptable AUC score.

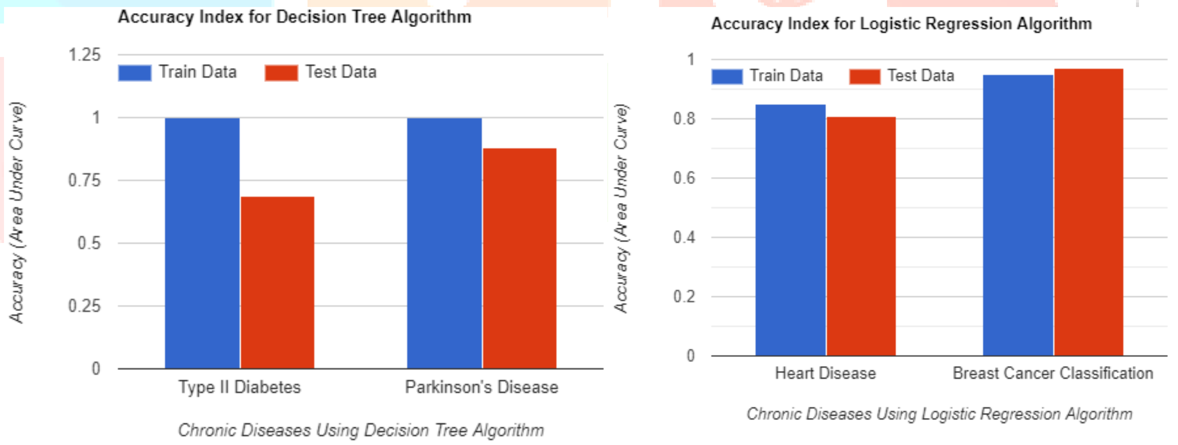

Figure 5. Comparison of AUC (Both Train and Test Data) for LR and DT

VIII. CONCLUSION AND ANALYSIS

In our undertaking, we deployed and assessed two machine learning algorithms, Logistic Regression and Decision Tree Classifiers, to forecast and identify Chronic Diseases.

Our findings reveal that the suggested technique attains superior performance on our dataset, presenting remarkable accuracy and robustness to changes. This research paper outlines the issue at hand, review of related literature, the prerequisites of the system, the dataset employed. Additionally, we elucidated on the training process, evaluation metrics, and comparison with other methods in the results segment. Our machine learning models were able to attain a respectable Mean Average Precision (mAP) of 0.823 on the training dataset and 0.767 on the test dataset. These outcomes demonstrate that the model is relatively successful in identifying chronic diseases.

The limitation with this research and project is the use of advanced data such X-Ray scans, Computerized Tomography (CT) Scans, Ultrasound Scans, etc., and with use of advanced data the Machine Learning models should also be more complex that can even cover other Chronic Diseases such as tumour or cancer, complex heart conditions, meningitis, Alzheimer's Disease, etc., There is also scope in the improvement of Accuracy and Robustness of the models.

REFERENCES

[1] Reddy, K. S., Shah, B., Varghese, C., and Ramadoss, A. (2005). “Responding to the threat of chronic diseases in India.” *The Lancet*, 366(9498), 1744–1749.

[2] Joshi, R., Cardona, M., Iyengar, S., Sukumar, A., Raju, C. R., Raju, K. R., Raju, K., Reddy, K. S., Lopez, A., and Neal, B. (2006). “Chronic diseases now a leading cause of death in rural India—mortality data from the Andhra Pradesh rural health initiative.” *International journal of epidemiology*, 35(6), 1522–1529.

[3] Alanazi, R. (2022). “Identification and prediction of chronic diseases using machine learning approach.” *Journal of Healthcare Engineering*, 2022.

[4] Chen, M., Hao, Y., Hwang, K., Wang, L., and Wang, L. (2017). “Disease prediction by machine learning over big data from healthcare communities.” *Ieee Access*, 5, 8869–8879.

[5] Lu, H., Uddin, S., Hajati, F., Moni, M. A., and Khushi, M. (2022). “A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus.” *Applied Intelligence*, 52(3), 2411–2422.

[6] Sinha, P., Sinha, P., et al. (2015). “Comparative study of chronic kidney disease prediction using KNN and SVM.” *International Journal of Engineering Research and Technology*, 4(12), 608–12.

[7] Battineni, G., Sagaro, G. G., Chinatalapudi, N., and Amenta, F. (2020). “Applications of machine learning predictive models in the chronic disease diagnosis.” *Journal of personalized medicine*, 10(2), 21.

[8] Priya, R. and Jinny, S. V. (2021). “Elderly healthcare system for chronic ailments using machine learning techniques—a review.” *Iraqi Journal of Science*, 62(9), 3138–3151.