International Journal Of Engineering Research Excellence And Applied Science
Volume-III (Issue 2) –June-2025 ISSN: 3048-5355

Early Disease Detection and Prediction using AI Technologies:

Arpit Singh[1], Ayush Sharma [2]
Department Of Computer Science & Engineering
Arya institute of Engineering and Technology Jaipur , Rajasthan India

Abstract: This paper offers an extensive and perceptive analysis of the present state of healthcare prediction. It underscores the significant benefits that have arisen from the integration of artificial intelligence, emphasizing its positive impact. The utilization of AI in healthcare prediction has brought significant advancements, but it also comes with its own set of challenges. This article aims to contribute to the advancement of disease detection and prediction by presenting the findings of an in-depth literature review encompassing recent research articles in the field. It also explores the potential impact of these findings. HealthCare prediction has become crucial for saving lives, and intelligent systems have emerged to analyse complex data relationships and generate valuable information for predictions. The paper reviewed many working papers and provided insights into the methodologies employed in each study. Additionally, it acknowledges the challenges that must be addressed to maximize the potential of artificial intelligence in disease diagnosis and prediction, and also it suggests the solution for challenges. Research has demonstrated that AI plays a significant role in accurate disease diagnosis, healthcare anticipation, and analysis of health data by leveraging large-scale clinical records and reconstructing patients' medical histories..

Keywords: *Machine Learning, Deep Learning, CNN, AI, RF*

1. Introduction

In today's world, people encounter a wide range of diseases as a result of their current environmental conditions and lifestyle choices. The early identification and prediction of these illnesses are of utmost importance to prevent their severity. [1,2] As per medical reports, the mortality rate among humans rises due to chronic diseases. Some of the prevalent chronic illnesses include diabetes, cardiovascular diseases, cancer, strokes, hepatitis C, and arthritis. Due to their prolonged duration and significant mortality rates, the accurate diagnosis of these conditions holds paramount importance in the healthcare sector Therefore, it is crucial to mitigate the factors contributing to a patient's risk of mortality. [3,4] The progress in medical research simplifies the collection of health-related data. Machine learning can streamline the analysis of patient data and other relevant information, contributing to the early detection of diseases. In the field of machine learning, a wide range of techniques is available, encompassing semi-supervised learning, supervised learning, unsupervised learning and deep learning. [5,6] To address this need, it is essential to create a machine learning model capable of taking input symptoms and forecasting the probability and risk of disease progression or its impact on an individual's well- being. The primary aim of is to utilize a machine learning approach for the identification and prediction of chronic diseases in individuals. The dataset again which is important part of process consists of two distinct types of information. First, it contains structured data encompassing details like the patient's age, gender, height, weight, and more [7]. This structured data deliberately excludes any personal identifiers such as the patient's name or ID.

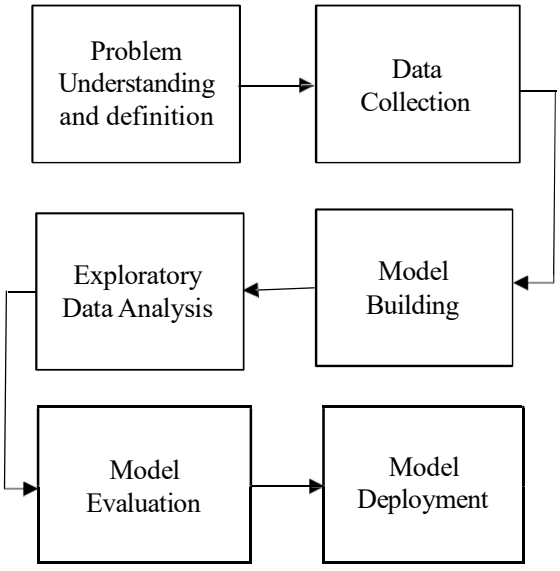

Fig1: Predictive Analytics Steps [8]

Second, the dataset also includes unstructured data comprising the patient's symptoms, records of

consultations with healthcare professionals about their condition, and information about their lifestyle.Fig.1 shows the Predictive analytics steps. The primary objective of this review is to succinctly and clearly explore the work done till date, technologies used and datasets used in medical diagnosis. This paper is structured as follows: Section 2 provides an overview of the existing research related to this study. In Section 3, we present the fundamentals and details of the algorithms employed in Disease Prediction and detection, the basics of datasets used for disease detection and prediction, the results parameters and a comprehensive discussion. Finally, we conclude our study. The list of references used in this research is included at the end of the paper

2. Review Of Literature

Viktor et al. (2021) [9], focuses on complications of skin of diabetes mellitus. The researchers utilized hyperspectral imaging and ANN techniques for a feel of real time image processing. sensitivity and specificity of the method is 95% and 85%.The authors mention a limitation in terms of the time it takes for hypercube acquisition and data transfer via USB port, which can be improved in future research.It can be improved by using Bluetooth or

wifi module in IOT for taking input instead of USB Port.**Ritesh Jha et al. (2022) [10]** conducted research on thyroid disease prediction. They employed techniques such as PCA, DT, KNN and

NN. The study aimed to provide solutions for predicting thyroid diseases. reduced dimension data is obtained by Dimension reduction was inputted into classifiers. To generate sufficient data, Data augmentation has been used. Complex diseases that pose a threat to life can be predicted using deep learning models.The accuracy is 99.95%, which is excellent in comparison to currently used methods. Can be worked on diabetes complication prediction using same tools and technologies.In the work by **Victor Chang et al. (2022) [11],** a RF classifier algorithm was improved for identifying disease of heart. The authors suggest that future research can explore invasive-based approaches and consider angiography as well. The accuracy of the developed algorithm was reported as 83%, and the authors acknowledge the potential for improvement in this aspect. Random forest along with ensembling can be used for better results.

Shahid Mohammad Ganie et al. (2022) [12] focused on predicting diabetes based on lifestyle indicators. They proposed a new hybrid based framework using lifestyle indicators for an early diagnosis of type 2 diabetes. Various ensemble learning techniques, such as voting, boosting, bagging were employed. The study incorporated performance measurement metrics and utilized techniques like SMOTE, oversampling, and k-fold cross-validation.The bagged decision tree had the greatest accuracy percentage of all the classification methods (99.41%).

To find probability of disease in patients and early Prediction of diseases can be future scope of the work.Above future scope can be met by using proper ML and DL algorithm"s ensembling.In a study by **V. Jackins et al. (2021)[13],** artificial intelligence techniques, including Naive Bayes classification and RF classification algorithms, were used to classify disease datasets for multiple diseases. The research showed that the RF model performed the best compared to remaining models. However, the authors noted that the accuracy of the model, which was reported as 74%, could be further improved, especially for real-time data.

Above future scope can be met by using ensembling Random forest classfier and IOT can be used for run time data. **Haohui Lu et al. (2021) [14]** a dataset of

type 2 diabetic mellitus (T2DM) in the actual world was used to create a collection of patient networks and machine learning techniques for disease prediction. 1,028 patients with T2DM and 1,028 individuals without T2DM were included in the dataset. To predict T2DM risk, eight ML models were used, including KNN, logistic regression, DT, XGBoost, SVM, naive Bayes, ,RF and ANN. Features including of the network closeness centrality, eigenvector centrality, and age of patient were shown to be the most crucial in the random forest model, which performed better. The study also emphasised the need for enhanced databases that include complete disease codes, standardised formats, and consistent data recordings. The thorough studies demonstrate that the performance of the suggested framework using machine learning classifiers ranges from 0.79 to 0.91 for AUC.For better accuracy, more proper dataset can be used.Accuracy can be improved by ensembling of algorithms and proper dataset can be obtained by using IOT technology with good values of confusion matrix. **Hamza Mustafa et al. (2022) [15]**

proposed an approach that combines deep neural networks and PCA to learn variations in raw image features for diabetic retinopathy detection. A machine learning ensemble classifier was employed to gain robust performance and high classification accuracy. Using Messidor-2 and EyePACS datasets with various numbers of categories, the performance of the system approach was compared to traditional CNN based approaches. The experimental results demonstrated superior performance, with accuracy reaching up to 95.58%. The study suggests that the proposed approach shows promise for automatic diabetic retinopathy detection, and the accuracy of the method was observed to increase with a decrease in the number of categories. 2 diabetic mellitus (T2DM) regulatory claim dataset to construct an outfit of understanding systems and machine learning strategies for ailment expectation. 1,028 patients with T2DM and 1,028 people without T2DM were included within the dataset. To anticipate the hazard of T2DM,8 ML models were utilized, counting calculated relapse, KNN, SVM, credulous Bayes, and some more. The eigenvector centrality, nearness centrality, and understanding age were shown to be the foremost significant components of the arbitrary woodland show, which beated other models. To consider moreover accentuated they require for upgraded databases that incorporate total malady codes, institutionalized groups, and reliable information recordings.

IOT can be combined with the above approach for better performance. **Nada Y. Philip et al. (2021)[16]** We suggested some tools for looking at information to help with problems caused by type 2 diabetes. This helps doctors and researchers see relationships between a patient's physical signs and the problems caused by their Type 2 Diabetes. The package contains predictive, exploratory and visual analytics providing features including patient‟s multi-tier profile classification for T2D, risk prediction for complications connected to T2D, and patient response prediction for certain medications.Precision value the authors got was 73.3%.

Future development opportunities include incorporating artificial intelligence techniques for more robust prediction models, perform clinical data analytics validation and train on larger databases to improve prediction accuracy. As a future scope Decision tree or random forest can be used for improvement. **Nikos Fazakis et al. (2021)[17]** They made a tool to guess if someone might get diabetes. They used parts of a process called Knowledge Discovery in Database. The research was about how to make a set of information, pick out important parts, and use computer programs to classify it. They came up with a computer program that predicts diabetes very well, with a score of 0. 884.The writers recommended improving the data by filling in missing information through techniques like IRSSI, and trying out additional ways to select important features.**Dritsas and Trigka (2022)[18]** emphasized the importance of early detection of diabetes syndrome, which is defined by shifts in carbohydrate, lipid, and protein metabolism. They discussed the use of supervised learning techniques to develop risk prediction tools for Type 2 diabetes mellitus (T2DM) with high efficacy. The study revealed that KNN and RF models performed the best among the compared models.The RF and KNN shown good results after using SMOTE with 10-fold cross-validation, with an accuracy of 98.59%.Further CNN and LSTM algorithms will be used on the same dataset and then compared with other relevant published studies in terms of their accuracy to extend the machine-learning framework.**Lu et al. (2022) [19]** highlighted the increasing level of chronic disorders like T2DM, which has placed a significant burden on healthcare systems. They created a collection of patient networks and machine learning methods for disease diagnosis by making use of a T2DM organizational claim dataset from the real world. The study came to the conclusion that the RF model's accuracy was superior to other models' accuracy.

In the future, huge amounts of more sophisticated and relevant CKD data will be gathered to assess disease severity and enhance the model performance.**Dong et al. (2022)[20]** The way that end-stage renal illness, cardiovascular sickness (CVD), and grimness in individuals with diabetes are fundamentally brought about by diabetic kidney condition. They developed prediction models using 46 medical features extracted from Electronic Medical Records (EMR) and applied seven different Machine Learning (ML) methods. The Light Gradient Boosting Machine (GBM) framework had the highest AUC, with a value of 0.815., indicating its effectiveness in predicting diabetic kidney syndrome.Further testing of the proposed model with a large dataset of up to millions of records and zero missing values is planned in the future, achieving an overall accuracy of 99.99%. **Aggarwal et al. (2022)[21]** investigated the susceptibility of diabetic individuals to coronavirus and developed a coronavirus risk forecasting model using a fuzzy inference framework and ML methods. While there is no evidence supporting a higher likelihood of infection in diabetic patients, the study aimed to address the higher

mortality rate associated with coronavirus in this population. The CatBoost classifier demonstrated the highest accuracy of 76% among all the classifiers considered.

In the future, a more effective method of generating synthetic data will make hyper-parameter optimization unnecessary by doing away with the inherent bias that comes from being entirely naïve and eliminating variance fluctuations.emphasized the importance of disease prediction and early detection for disease prevention. They employed Support Vector Machine, Artificial Neural Network.**Ahmed et al. (2022)[22]** methods in a fused ML technique to improve disease diagnosis. The proposed combined ML framework achieved a predicted accuracy of 94.87%, surpassing previously reported techniques.Future models can be gathered using a cloud storage system. The fused model evaluates if a patient has diabetes based on their most recent medical data.**Singh et al. (2022)[23]** focused on chronic kidney disease and presented a deep-learning framework for earlyidentification and prediction of the disease. Their deep neural network model outperformed other ML strategies, achieving a perfect accuracy rate of 100%.

Helalay et al., (2022) [24] demonstrated that AD is a long-lasting and irreversible brain disorder; there is currently no treatment that can effectively treat it. However, the currently available treatments can slow the progression of the disease. As a result, the prior detection of AD is an extremely important factor in precluding and controlling the progression of the disease. Both Convolutional Neural Networks (CNN) and VGG19 were utilized in this study as classification strategies for medical images to identify AD. In the end, it was determined that the VGG19 pre-trained framework performed better than the CNN and accomplished an accuracy of 97% for the classification of multi-class AD stage data.In the future next variant of VGG19 can be used. **Lamba et al., (2022) [25]** Parkinson's disease is a neurodegenerative syndrome that moves through its stages slowly. Because its symptoms develop for the disease, it can be difficult to diagnose it in its early stages. The authors of this study postulate a speech signal-based composite Parkinson's disease diagnosis system as a means of performing an early assessment of the condition. The speech dataset was utilized to perform performance analysis on the various combination possibilities. In the end, it was determined that the best performance was achieved by combining the Genetic Algorithm (GA) and the RF classifier. This combination achieved a precision of 95.58%. Table 1 depicts the comparison of the reviewed literature of various authors. Better performance can be achived by combining the Genetic Algorithm (GA) and the RF classifier and SVM. **Michele Bernardini et al. (2021) [26]** authors are using electronic health records to predict the chances of getting an eye problem called Diabetic Retinopathy, which can happen to people who have diabetes. We want to know when someone is most likely to get this problem. They made a new way to prepare data and gave a collection of information from different places about diabetes that has been marked and organized.Area Under the Precision-Recall Curve of, respectively, 72.43% and 84.38%

In the future, may try to predict other problems that diabetics can have, like heart disease, kidney problems, nerve problems, and blood vessel problems. XGB, LR, DT, RF, SVM, NB algorithms are compared. Future work is to predict other diseases due to diabetes.can be done with proper dataset and ML and DL algortihms. **Mohamed M. Farag et al. (2022)[27]** authors proposed a novel idea for automatically determining the critical effects of Diabetic Retinopathy from a single Colour Fundus Photograph (CFP) using deep learning techniques. Author‟s method leverages the Convolu- tional Block Attention Module (CBAM) to enhance the model's discriminative capability. Additionally, we employ the visual embedding extracted from DenseNet169's encoder to further improve performance. The model is trained on dataset, obtained from Kaggle. Author approach demonstrates promising results, outperforming existing methods on the with an impressive 97% accuracy. **Nahla H. Barakat et.ai.,(2010)[28]** Worked on diabetes diagnosis using SVM and got the accuracy of 94%.As a future scope, can be worked on diabetes complications using various ML algorithms. **Min Chen et al. (2017)[29]** aimed to predict chronic disease outbreaks in disease- frequent communities. They streamlined machine learning algorithms and experimented with new prediction models using central China collected real-life hospital data. latent factor model is used to fill in the gaps in the data to deal with incomplete data. Their research focused on a cerebral infarction

i.e. regional chronic disease. New CNN is used. Accuracy is 94.8%,which can be further enhanced by