International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Early Disease Detection and Prediction using AI Technologies: Approaches, Future Outlook, Mitigation Strategies, and Synthesis of Systematic Reviews

Anita Dombale[1], Premanand Ghadekar [2]

Submitted: 28/01/2024 **Revised:** 06/03/2024 **Accepted:** 14/03/2024

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.

1Dept. of Computer Engineering, Vishwakarma Institute of Technology (Savitribai Phule Pune University), Pune, India. bhanuseanita@gmail.com
2Dept. of Information Technology, Vishwakarma Institute of Technology (Savitribai Phule Pune University), Pune, India. premanand.ghadekar@vit.edu

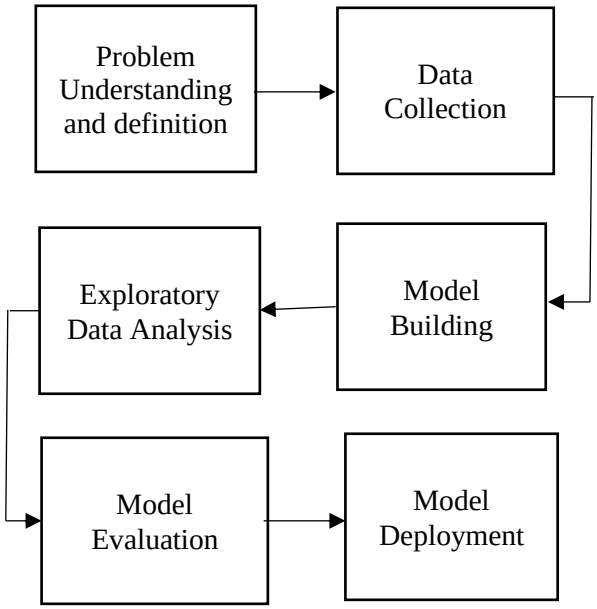

Fig1: Predictive Analytics Steps [8]

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 early identification 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 Convolutional 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 using improved CNN algorithm.**Tawfik Beghriche et al. (2021)[30]** presented a Deep Neural Network (DNN)-based effective medical decision-making system for diabetes prediction. They highlighted the effectiveness of DNN algorithms in various domains and emphasized their potential for prediction and diagnosis purposes in healthcare.Accuracy of 99.75% is obtained which is far better as compared to existing results. Further diabetic complications can be predicted using ML and DL algorithms or combinations of both. **Divyashree N. et al. (2022)[31]** It was advised to design and build a web-based clinical decision support .Their solution aims to make CDSS usable on desktops and mobile devices for both regular people and physicians. To forecast coronary artery disease, it combined predictive analytics with the LWGMK-NN algorithm and prescriptive analytics with prescription rules. By incorporating further characteristics like the patient's medical history and current drugs. future work may include the incorporation of automatically personalized medication prescriptions. this can be done using datasets and ML and DL algorithms.**Yunlei Sun et al. (2019)[32]** Authors studied information about diabetes in patients who were in the hospital, used electronic records to learn about their diabetes diagnosis, how much sugar was in their blood, and other tests related to diabetes. The goal was to use computer techniques to study diabetes.This research suggests using a Convolutional Neural Network to make a model for diagnosing diabetes in medical settings. They suggest combining this method with another

technique called BN layer. The CNN method helps with applying convolution to one-dimensional datasets that don't have a connection. The BN layer helps to better train the model, make it faster, and more accurate.

The accuracy obtained for training data is 99.85%. Limitation is data used in this study is limited. Therefore, future research should focus on improving the model's performance when dealing with high-dimensional and large-scale data. Additionally, optimizing and enhancing the efficiency of the proposed model is another target for future investigations. The research design model could also be extended to other one-dimensional unrelated datasets, such as other electronic medical records information in the medical domain. efficiency of the proposed model can be enhanced by using ensembling techniques.

3. Technologies Used

3.1 Machine Learning

Machine learning is a subset of artificial intelligence (AI) that entails the process of training algorithms using data, enabling them to make predictions or execute actions without the need for explicit programming. Machine Learning involves many algorithms which comes under 2 types i.e. Supervised Learning and Unsupervised Learning. Supervised machine Learning is of two types: a) Classification b) Regression

3.1.1 Supervised learning classification algorithms:

The types of classification Supervised learning algorithms are decision trees, support vector machines, naïve Bayes, K-nearest neighbours, and neural networks.

- **Decision Tree**

The Decision Tree is used for each internal node to represent a feature tree, a leaf node to represent a class label, and branches to indicate conjunctions of features. In a non-parametric supervised learning approach, decision trees are used in both regression and classification. Calculation of the decision tree is done using information gain and entropy. Equation 1 provides the equation for calculating the entropy:

$$E(S) = \sum_{i=1}^{c} -p_i log_2 p_i \qquad (1)$$

Where, p_i is the probability of an element/class 'i' of the data.

Equation 2 displays the formula for determining the information gain

Gain(T,X) = Entropy(T)-Entropy(T,X) (2)

Where T is the target value and X is the actual variable of the dataset [33]

which disease is positive and which is not can be categorized through this Decision tree algorithm.

- **SVM**

It combines supervised regression and classification learning methods. As a result, a higher dividing hyperplane can be constructed by projecting the input vector to a higher-dimensional space. [34]. It utilizes the dividing or separating hyperplane with the expression to view this training information:

$$w.m + b=0 \qquad (3)$$

In this equation 3, b is a scalar, and m is a dimensional vector, w is perpendicular to the horizontal dividing hyperplane. As seen in Figure 2, an SVM trained on instances from 2 has maximum edge hyperplane classes [35].

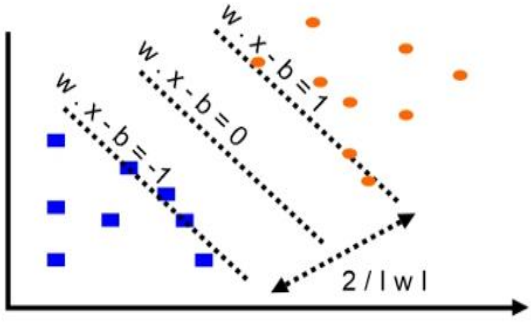

Fig 2. An SVM trained with samples from 2 classes.[35]

- **NAIVE BAYES**

Naive Bayes classifiers encompass a set of classification techniques rooted in Bayes' Theorem. They constitute a family of algorithms characterized by a common principle: the independence assumption, meaning that every pair of features being classified is treated as independent of each other

- **KNN**

The K-NN algorithm operates on the assumption of similarity between the new data point and existing cases, assigning the new data point to the category that bears the closest resemblance among the available categories.

- **RANDOM FOREST**

The Random Forest (RF) classifier, the better it is at being accurate. It makes a bunch of trees called a forest that work together to predict things better. Every decision tree in the random forest is made using only part of the data and trained using estimates. The RF algorithm tries to make the best choice by putting together the results from many DTs.

With the help of this algorithm, results can be obtained in the form of tree and disease prediction can be done in a better way.

3.1.2 Supervised learning Regression algorithms

Regression is a crucial and widely employed statistical and machine learning technique. Its primary goal is to forecast continuous numeric output labels or responses based on input data. The model's predictions are derived from the knowledge acquired during the training phase.

Some of Types of Regression Algorithms are Simple linear Regression, Logistic Regression, Ensemble Method.

3.1.3 Unsupervised machine learning algorithms:

Unsupervised learning is a machine learning approach where models are not guided by a training dataset. Instead, these models autonomously discover concealed patterns and insights within the provided data.

Here is a compilation of well-known unsupervised learning algorithms

- **K-means clustering**

This iterative algorithm partitions an unlabelled dataset into k distinct clusters, ensuring that each

data point belongs to a single group with similar properties

- **Hierarchical clustering**
  Hierarchical clustering is an algorithm that constructs a hierarchical structure of clusters from a given dataset. Initially, each data point is considered as its own cluster. As the algorithm progresses, it merges clusters that are close to each other based on a similarity or distance metric. This process continues until there is only one cluster remaining in the hierarchy. This approach allows for the creation of a tree-like structure, known as a dendrogram, which visually represents the relationships between clusters at different levels of granularity.

- **K-Nearest Neighbours**
  K-Nearest Neighbours (KNN) is a straightforward algorithm that retains all available data points and classifies new instances by assessing their similarity to existing cases. It performs effectively when there is a measurable distance between data points. However, its computational efficiency tends to decrease when working with extensive training sets due to the need to calculate distances between data points, which can be nontrivial.

- **Principle Component Analysis**
  This technique is used to perform data reduction or dimensionality reduction. Using PCA, the most important events are identified and monitored throughout the experiment [36,37].

3.2 Deep Learning Algorithms

Deep learning algorithms play a critical role in feature extraction and processing for a wide range of data types, including structured and unstructured data. Nevertheless, it's worth acknowledging that these algorithms may not always be the most suitable choice for tasks that involve intricate problems, as they typically demand access to extensive datasets to operate optimally**.**

Here is a compilation of well-known deep learning algorithms

- **Convolutional Neural Networks (CNNs)**
  CNN is also recognized as ConvNet, and it is a category of Artificial Neural Network(ANN) with a deep feed-forward construction and incredible simplifying capability when associated with more networks with fully connected layers [38]. Figure 6 depicts CNN's core conceptual paradigm.

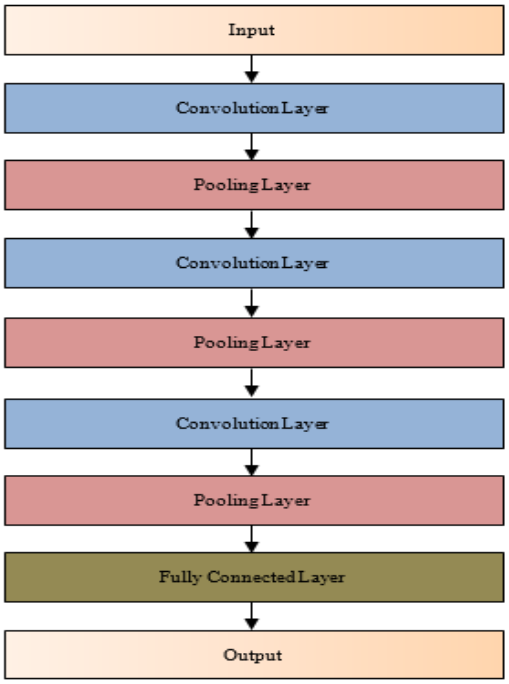

Fig 3. CNN [38]

- **Long Short-Term Memory Networks (LSTMs)**
  Long Short-Term Memory networks, can be characterized as a type of Recurrent Neural Network (RNN) designed to acquire and adapt to long-term dependencies in data. They excel at retaining and retrieving past information over extended periods, which is their primary function.

  LSTMs are specifically engineered to retain information over time, making them particularly valuable in time series prediction tasks where the ability to capture and preserve memory of past inputs is crucial.

- **Generative Adversarial Networks (GANs)**

  GANs, are deep learning algorithms utilized for the generation of new data instances that closely resemble the training data. GANs typically consist of two key components: a generator, which learns to create synthetic data, and a discriminator, which refines its discriminatory abilities by learning from this synthetic data. Over time, GANs have gained significant popularity, finding extensive application in tasks such as enhancing astronomical images and simulating gravitational lensing caused by dark matter**.**

- **DATASETS**
  Healthcare datasets pose a formidable challenge for analysis owing to their vast scale and intricate nature. Healthcare datasets encompass various types

of information sources, with some notable examples being electronic health records (EHRs), which serve as digital repositories for comprehensive patient medical data. Another common type is claims datasets, which furnish insights into the healthcare services received and the corresponding expenses incurred. In addition to these, there exist disease registries, housing data pertinent to individuals afflicted with particular diseases or conditions, and clinical trial datasets, which encompass details regarding trial participants, interventions administered, and the outcomes observed.

Datatypes can be of any type like MRI readouts, sonography, Social media claimed data, Electronic health records, Behavioural data, clinical data. Many datasets are readily available on many websites, which can also be used.

- **Feature Extraction**
  Feature engineering is the practice of transforming a dataset to enhance the performance of a machine learning model during training. It is Modifying the dataset through actions such as adding, removing, merging, or altering features. This meticulous adjustment of the training data is done with the aim of ensuring that the resultant machine learning model is well-suited to meet its objectives. A variety of techniques exist for feature extraction, such as principal component analysis (PCA), autoencoders, filter methods, wrapper methods etc.

Following Table1 shows the work done on disease detection and Prediction

Results

- **Evaluation Matrix of Supervised Classification Algorithms**
  The evaluation of supervised classification algorithms often involves assessing their performance using metrics such as accuracy, sensitivity, and specificity. These metrics provide insights into how well the model is performing in different aspects of classification.[39]

- Accuracy: Accuracy is a metric used to gauge the overall correctness of a model's predictions. To calculate accuracy, we divide the sum of correctly predicted instances, which includes both true positives and true negatives, by the total number of instances in the dataset.
- Sensitivity (True Positive Rate or Recall): Sensitivity gauges the model's ability to correctly identify positive instances. It is calculated as the ratio of true positives to the total number of actual positive instances. The formula for sensitivity is:

  Sensitivity = TP / (FN + TP)

- Specificity (True Negative Rate): Specificity assesses the model's capability to correctly identify negative instances. It is determined by the ratio of true negatives to the total number of actual negative instances. Specificity can be expressed as:

  Specificity = TN / (FP + TN)

- F1 Score: It is calculated as:

  F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

- **Evaluation Matrix of Supervised Regression Algorithms**:
  Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are frequently employed metrics for assessing the effectiveness of regression models.

- **Evaluation Matrix of Unsupervised Clustering Algorithms**
  Evaluating unsupervised clustering algorithms can be a bit challenging since there are no predefined class labels to compare the results against, as is the case with supervised learning.

  Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index (Variance Ratio Criterion), Dunn Index, Adjusted Rand Index (ARI) are some of commonly used evaluation metrics.

Discussion

Supervised and unsupervised machine learning methods and Deep Learning methods have demonstrated considerable promise in healthcare applications. All approaches possess distinct strengths and weaknesses, and their appropriateness for healthcare tasks hinges on the data's characteristics and the specific goals of the analysis.

It's important to note that while deep learning holds great promise, there are challenges in deploying the models in healthcare. Data privacy and security are paramount concerns, and ensuring the reliability and interpretability of deep learning models in critical medical decision-making remains an ongoing research area.

Conclusion

In conclusion, predictive health refers to the use of predictive analytics and ML and DL algorithms to improve health and healthcare services. The adoption of machine learning and deep learning

techniques has the potential to revolutionize traditional healthcare delivery. Healthcare data is recognized as a crucial component that contributes to the advancement of medical-care systems. The availability of diverse sources of health data has increased tremendously in current years.

The history of a predictive analytics tool, its field of use, and its approach for predicting Disease have all been covered in this paper.

The paper discussed the background of a Predictive Analytics Tool and its domain, focusing on the methodology for early disease prediction. It emphasizes the potential of artificial intelligence (AI) in enhancing the quality of work in healthcare. The paper reviewed many working papers and provided insights into the methodologies employed in each study. This paper also finds out limitations of studied research paper and suggests possible solution for it. 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. However, further studies are needed to improve AI integration with healthcare data quality management considerations.

References

[1] Alam, Talha Mahboob, Muhammad Atif Iqbal, Yasir Ali, Abdul Wahab, Safdar Ijaz, Talha Imtiaz Baig, Ayaz Hussain, et al. "A model for early prediction of diabetes." Informatics in Medicine Unlocked 16 (2019): 100204.

[2] MDPI and ACS Style An, Q.; Rahman, S.; Zhou, J.; Kang, J.J. A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges. Sensors 2023, 23, 4178. https://doi.org/10.3390/s23094178

[3] Arroba, Ana I., and Manuel Aguilar-Diosdado. "Special Issue “The Prevention, Treatment, and Complications of Diabetes Mellitus”." Journal of Clinical Medicine 11, no. 18 (2022): 5305.

[4] Rayan Alanazi, "Identification and Prediction of Chronic Diseases Using Machine Learning Approach", Journal of Healthcare Engineering, vol. 2022, Article ID 2826127, 9 pages, 2022. https://doi.org/10.1155/2022/2826127

Author Contributions:

Conceptualization, A.D. and P.G.; resources, P.G.; data curation, A.D. and P.G.; writing-original draft preparation, A.D.;

Acknowledgements: The authors would like to extend their appreciation to the anonymous reviewers for their valuable and insightful feedback, which significantly contributed to the refinement of the paper in its current form.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest

Declarations

Ethics approval The authors declare that this review was completed in compliance with ethical standards.

Consent to participate Not applicable

Consent for publication Not applicable

[5] MDPI and ACS Style Bakator, M.; Radosav, D. Deep Learning and Medical Diagnosis: A Review of Literature. Multimodal Technol. Interact. 2018, 2, 47.

[6] Dritsas E, Trigka M. Data-Driven Machine-Learning Methods for Diabetes Risk Prediction. Sensors (Basel). 2022 Jul 15;22(14):5304. doi: 10.3390/s22145304. PMID: 35890983; PMCID: PMC9318204.

[7] Salmond, Susan W., and Mercedes Echevarria. "Healthcare transformation and changing roles for nursing." Orthopedic nursing 36, no. 1 (2017): 12.

[8] https://www.analyticsvidhya.com/blog/2022/09/the-6-steps-of-predictive-analytics/

[9] Viktor Dremin , Zbignevs Marcinkevics, Evgeny Zherebtsov , Alexey Popov, Andris Grabovskis,

[10] Hedviga Kronberga, Kristine Geldnere, Alexander Doronin, Igor Meglinski , Senior Member, IEEE,and Alexander Bykov. Skin Complications of Diabetes Mellitus Revealed by Polarized HyperspectralImaging and Machine

Learning IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 40, NO. 4, APRIL 2021

[11] Ri Ritesh Jha1 · Vandana Bhattacharjee1 · Abhijit Mustafi1 Increasing the Prediction Accuracy for Thyroid Disease:A Step Towards Better Health for Society Wireless Personal Communications (2022) 122:1921–1938https://doi.org/10.1007/s11277-021-08974-3

[12] Victor Chang et al., An artificial Inteeligence model for heart disease detection using machine learning algorithms,Healthcare Analytics

[13] Shahid Mohammad, Majid Bashir Malik., An Ensamble machine learning Approach for predicting Type-II diabetes mellitus based on lifestyle indicators.,Elsevier

[14] V. Jackins1 · S. Vimal1 · M. Kaliappan2 · Mi Young Lee3 AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes

[15] Haohui Lu et al.,A patient network based machine learing model foe disease Prediction:The case of Type 2 diabetes mellitus 2021

[16] Hamza Mustafa et al .,Multi-stream deep neural netoerk for Daibetic Retinopathy Severity classification under a boosting Framwork,IEEE Accesss 2022

[17] Nada Y. Philip et al.,(2021) A Data Analytics suite for Exploratory Predictive,and visual Analysis of Type 2 Daibetes.

[18] Nikos Fazakis et al.,Machine Learing Tools for Long term type 2 Diabetes Risk Prediction

[19] Dritsas, Elias, and Maria Trigka. "Data-driven machine-learning methods for diabetes risk prediction." Sensors 22, no. 14 (2022): 5304.

[20] Lu, Haohui, Shahadat Uddin, Farshid Hajati, Mohammad Ali Moni, and Matloob Khushi. "A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus." Applied Intelligence 52, no. 3 (2022): 2411-2422.

[21] Dong, Zheyi, Qian Wang, Yujing Ke, Weiguang Zhang, Quan Hong, Chao Liu, Xiaomin Liu, et al. "Prediction of 3-year risk of diabetic kidney disease using machine learning based on electronic medical records." Journal of Translational Medicine 20, no. 1 (2022): 1-10.

[22] Aggarwal, Alok, Madam Chakradar, Manpreet Singh Bhatia, Manoj Kumar, Thompson Stephan, Sachin Kumar Gupta, S. H. Alsamhi, and Hatem Al-Dois. "COVID-19 Risk Prediction for Diabetic Patients Using Fuzzy Inference System and Machine Learning Approaches." Journal of Healthcare Engineering 2022 (2022).

[23] Ahmed, Usama, Ghassan F. Issa, Muhammad Adnan Khan, Shabib Aftab, Muhammad Farhan Khan, Raed AT Said, Taher M. Ghazal, and Munir Ahmad. "Prediction of diabetes empowered with fused machine learning." IEEE Access 10 (2022): 8529-8538.

[24] Singh, Vijendra, Vijayan K. Asari, and Rajkumar Rajasekaran. "A deep neural network for early detection and prediction of chronic kidney disease." Diagnostics 12.1 (2022):

[25] .Helalay et al., (2022),Paper on brain disorder

[26] Lamba, Rohit, et al. "A systematic approach to diagnose Parkinson's disease through kinematic features extracted from handwritten drawings." *Journal of Reliable Intelligent Environments* (2021): 1-10

[27] MICHELE BERNARDINI 1, LUCA ROMEO 1,2, ADRIANO MANCINI1, A Clinical Decision Support System to Stratify the Temporal Risk of Diabetic Retinopathy,2021

[28] AND EMANUELE FRONTONI 1, (Member, IEEE), Katzfuss, Matthias, Jonathan R. Stroud, and Christopher K. Wikle. "Understanding the ensemble Kalman filter." The American Statistician 70, no. 4 (2016): 350-357

[29] Automatic Severity Classification of Diabetic Retinopathy Based on DenseNet and Convolutional Block Attention Module MOHAMED M. FARAG 1, MARIAM FOUAD 1,2, AND AMR T. ABDEL-HAMID

[30] Barakat NH, Bradley AP, Barakat MN. Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE Trans Inf Technol Biomed. 2010 Jul;14(4):1114-20. doi: 10.1109/TITB.2009.2039485. Epub 2010 Jan 12. PMID: 20071261.

[31] Disease Prediction by Machine Learning Over Big Data From Healthcare Communities MIN CHEN1, (Senior Member, IEEE), YIXUE HAO1, KAI HWANG2, (Life Fellow, IEEE), LU WANG1,

AND LIN WANG3,41School of Computer Science and Technology, Huazhong1Department of Industrial and Systems Engineering, Dongguk University, Seoul 04620, South

[32] AnEfficient Prediction System for Diabetes Disease Based on Deep Neural Network Tawfik Beghriche ,1 Mohamed Djerioui ,2 Youcef Brik ,2 Bilal Attallah ,2 and Samir Brahim Belhaouari 3 Grooms, Ian. "A comparison of nonlinear extensions to the ensemble Kalman filter." Computational Geosciences (2022): 1-18.

[33] Design and Development of We-CDSS Using Django Framework: Conducing Predictive and Prescriptive Analytics for CoronaryArtery Disease DIVYASHREE N. , (Member, IEEE), AND NANDINI PRASAD K. S., (Senior Member, IEEE)

[34] Yunlei Sun et al. (2019) worked on diabetes

[35] Mokrani, Hocine, Razika Lounas, Mohamed Tahar Bennai, Dhai Eddine Salhi, and Rachid Djerbi. "Air quality monitoring using IoT: A survey." In 2019 IEEE International Conference on Smart Internet of Things (SmartIoT), pp. 127-134. IEEE, 2019.

[36] Sathyadevan, Shiju, and Remya R. Nair. "Comparative analysis of decision tree algorithms: ID3, C4. 5 and random forest." In Computational intelligence in data mining-volume 1, pp. 549-562. Springer, New Delhi, 2015.

[37] Karatsiolis, Savvas, and Christos N. Schizas. "Region-based Support Vector Machine algorithm for medical diagnosis on Pima Indian Diabetes dataset." In 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), pp. 139-144. IEEE, 2012.

[38] Wang, Wei, Mengxue Zhao, and Jigang Wang. "Effective android malware detection with a hybrid model based on deep autoencoder and convolutional neural network." Journal of Ambient Intelligence and Humanized Computing 10, no. 8 (2019): 3035-3043.

[39] Supreetha, B. S., Narayan Shenoy, and Prabhakar Nayak. "Lion algorithm-optimized long short-term memory network for groundwater level forecasting in Udupi District, India." Applied Computational Intelligence and Soft Computing 2020 (2020)

[40] Ghosh, A., Sufian, A., Sultana, F., Chakrabarti, A., & De, D. (2020). Fundamental concepts of convolutional neural network. In Recent Trends and Advances in Artificial Intelligence and Internet of Things (pp. 519-567). Springer, Cham.

[41] https://towardsdatascience.com/taking-the-confusion-out-of-confusion-matrices-c1ce054b3d3e

[42] Prediction of heart diseases using random forest M Pal, S Parija - Journal of Physics: Conference Series, 2021 - iopscience.iop.org

[43] COVID-19 patient health prediction using boosted random forest algorithm C Iwendi, AK Bashir, A Peshkar, R Sujatha in public health, 2020 - frontiersin.org

[44] Support vector machine deep mining of electronic medical records to predict the prognosis of severe acute myocardial infarction X Zhou, X Li, Z Zhang, Q Han, H Deng, Y Jian - Frontiers in, 2022 - frontiersin.org

[45] Application of support vector machine for prediction of medication adherence in heart failure patients YJ Son, HG Kim, EH Kim, S Choi Healthcare informatics, 2010 - synapse.koreamed.org

[46] Disease Prediction Based on Individual's Medical History Using CNN M Krishnamoorthy, MSA Hamee - 2021 20th IEEE, 2021 - ieeexplore.ieee.org

[47] Novel deep learning architecture for predicting heart disease using CNN S Hussain, SK Nanda, S Barigidad - 2021 19th OITS, 2021 - ieeexplore.ieee.org

[48] Purushottam, K. Saxena and R. Sharma, "Efficient heart disease prediction system using decision tree," International Conference on Computing, Communication & Automation, Greater Noida, India, 2015, pp. 72-77, doi: 10.1109/CCAA.2015.7148346.

[49] H. Hartatik, M. B. Tamam and A. Setyanto, "Prediction for Diagnosing Liver Disease in Patients using KNN and Naïve Bayes Algorithms," 2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS), Manado, Indonesia, 2020, pp. 1-5, doi: 10.1109/ICORIS50180.2020.9320797.

[50] M. Saw, T. Saxena, S. Kaithwas, R. Yadav and N. Lal, "Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning," 2020 International Conference on Computer Communication and Informatics

(ICCCI), Coimbatore, India, 2020, pp. 1-6, doi: 10.1109/ICCCI48352.2020.9104210.

[51] I. Yekkala, S. Dixit and M. A. Jabbar, "Prediction of heart disease using ensemble learning and Particle Swarm Optimization," 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon), Bengaluru, India, 2017, pp. 691-698, doi: 10.1109/SmartTechCon.2017.8358460.