[1]Dr Kannan Vishwanatth

[2]Dr Savitha Satish

[3]Dr Chandra Sekhar Mohapatra

[4]Mihir Bharatkumar Anjaria

[5]Shahnazeer C K

[6]Akshay Sharma

Predictive Analytics and Machine Learning in Disease Diagnosis: A Review of Recent Advances

Abstract: - Artificial intelligence (AI) refers to the utilisation of computer information to exhibit intelligent behaviour with minimal human intervention, whereas machine learning (ML) is seen as a subset of AI methodologies. Typically, this form of intelligence is widely recognised to have originated with the advent of robotics. Given the slow progression of diseases, it is crucial to make early predictions and administer appropriate medication. Hence, it is imperative to present a decision model that may aid in the diagnosis of chronic illnesses and forecast future patient prognoses. The primary objective of this study is to investigate the application of Predictive Analytics and ML in the field of Disease Diagnosis. The study will focus on reviewing the latest advancements in this area. This work specifically emphasises the significance of ML prediction models in disease diagnosis within the AI field, amidst other approaches available. The study utilises a qualitative research methodology. ML has been prominent in the medical field as it offers methods for analysing disease-related data, as indicated by this study. ML algorithms are crucial in attaining early disease detection. Another crucial finding in this research is that the accuracy and performance of the model can be enhanced by employing an alternative technique to generate a single ensemble model.

Keywords: ML; Predictive Analysis; Medical Field; Disease Diagnosis.

INTRODUCTION:

Every day, the nature of human existence undergoes changes, while the overall health of each generation either progresses or declines. Life is inherently characterised by perpetual uncertainties. Encounter folks with life-threatening health conditions on occasion as a result of delayed disease discovery. Regarding the adult population, chronic liver disease would impact over 50 million individuals globally [1]. Nevertheless, if the illness is detected at an early stage, it can be halted. ML can be used to forecast diseases, enabling the early identification of prevalent

[1] Postdoctoral Research, Haas School of Business, University of California Berkeley

Email Id: kannan.vishwanatth@berkeley.edu

[2]Department of Infectious Disease, Johns Hopkins University Bloomberg School of Public Health,

Baltimore, Maryland, USA

Email Id: savitasatish@yahoo.com

[3]Associate Professor, Department of Pathology, Sri Ram Chandra Bhanj Medical College & Hospital (SCBMCH), Cuttack, Odisha

Email Id: drcsmohapatra@gmail.com

[4]Proprietor, Shuchi Laboratory, Ph.D. Research Scholar, Sunrise University, Alwar, Rajasthan

Email Id: shuchilaboratory@gmail.com

[5]Research Scholar, Pondicherry University Karaikal Campus, Department of Computer Science

School of Engineering & Technology

[6]Researcher, Chandigarh University, Delhi

Email Id: opto19bop1018@gmail.com

ailments. Presently, the prioritisation of health has taken a backseat, resulting in a multitude of issues. A significant number of patients face financial constraints that prevent them from seeking medical attention, while others are burdened with hectic schedules. However, neglecting persistent symptoms for a prolonged period can lead to substantial health consequences [2].

Global diseases provide a significant challenge, prompting medical experts and researchers to make maximum efforts in order to decrease mortality rates associated with these diseases. Predictive analytic models have become crucial in the medical field due to the growing amount of healthcare data obtained from many sources that are different and incompatible with each other [3]. However, the task of managing, retaining, and examining the vast quantities of past data and real-time data generated by healthcare services has posed an unparalleled difficulty when relying on conventional database storage methods. The evidence consists of data obtained from the evaluation of a patient and substances produced by the patient. Illnesses are conceptual medical entities that identify abnormalities in the observed evidence [4].

Predictive analytics is a crucial necessity in the healthcare industry. The accuracy of illness prediction can have a substantial impact, potentially resulting in the preservation of patients' lives through early and precise predictions [5]. Therefore, it is crucial to have dependable and effective techniques for healthcare predictive analysis. This study intends to provide a thorough examination of the current ML and deep learning methods used in healthcare prediction and identify the inherent challenges in utilising these methods in the healthcare field.

LITERATURE REVIEW:

The following table provides details on recent works pertaining to the research topic of predictive analytics and ML in disease diagnosis.

Table 1: Related Works

AUTHORS AND YEARS	METHODOLOGY	FINDINGS
Battineni et al., (2020) [6]	This study searched PubMed (Medline) and CINAHL libraries for 453 papers published between 2015 and 2019. In the end, 22 papers were chosen to clearly describe CD diagnostic and utilisation models of various illnesses with their strengths and weaknesses.	Since each method has pros and cons, the results showed that there is no one ideal way in real-time clinical practice. SVM, LR, and clustering were the most popular approaches. These models are useful for CD classification and diagnosis and are predicted to grow in medical practice.
Kavitha et al., (2021) [7]	The suggested study used regression and classification on the Cleveland heart disease dataset. ML uses Random Forest and Decision Tree.	Experimental results demonstrated 88.7% accuracy for the hybrid heart disease prediction model. The interface uses a Decision Tree-Random Forest hybrid model to forecast heart disease from user input.
Ramesh et al., (2022) [8]	ML technologies were used to analyse vast volumes of complex healthcare data to help doctors forecast diseases. An online UCI dataset with 303 rows and 76 attributes was used in this study. Some 14 of these 76 attributes are tested to compare technique performance.	The experiments showed that KNN with eight neighbours outperformed “Naive Bayes, SVM (Linear Kernel), Decision Tree Classifier with 4 or 18 features, and Random Forest classifiers in terms of effectiveness, sensitivity, precision, and accuracy, F1-score”.
Arumugam et al., (2023) [9]	Qualitative research methodology was used	The decision tree model consistently outperformed the naive Bayes and support vector machine models. This

		study optimized the decision tree model to achieve the highest accuracy in predicting the probability of heart disease in patients with diabetes.
Alqahtani (2023) [10]	The suggested BDA-CSODL technique aims to accurately identify medical images and diagnose diseases. BDA-CSODL approach involves phases including pre-processing, segmentation, feature extraction, and classification.	Comprehensive simulations on benchmark medical image datasets show the superiority of the proposed BDACSODL approach across various criteria.
Caruccio et al., (2024) [11]	This study compared ChatGPT and standard ML models for diagnosing low- and medium-risk diseases solely by symptoms.	The trials used two medical datasets with over 100 symptoms associated with multiple diagnoses for disease prediction.

Research Gap: Prior research has indicated that predictive analytics are essential for the healthcare industry. The accuracy of disease prediction can have a profound effect, potentially saving lives when predictions are accurate and timely, but also posing a risk to patients' lives when predictions are erroneous. Therefore, it is imperative to make precise and reliable predictions and estimations of diseases. Therefore, there is a need for reliable and effective techniques for healthcare predictive analysis. The objective of this research is to provide a thorough examination of prevalent ML and deep learning methods used in healthcare prediction, while also highlighting the inherent challenges connected with using these approaches in the healthcare field.

METHOIDOLOGY:

This study utilised a Qualitative research methodology by facilitating a more comprehensive examination of the contextual elements, trends, and opportunities associated with these technologies. The secondary data obtained from recent scholarly papers, studies, and articles that discuss the application of AI & ML in the healthcare industry. This data provided insights into the current state of AI and ML in the sector, including emerging trends, challenges, and potential opportunities. The project gathered secondary data from a range of sources, such as academic databases like JSTOR, Science Direct, and Google Scholar, as well as specialised health sector publications and reports. In addition, a comprehensive search was carried out in prominent publications specialising in health, artificial intelligence, and ML, including the Journal of Health, IEEE Transactions on Neural Networks and Learning Systems, and the Journal of ML Research.

RESULTS AND DISCUSSIONS:

Disease diagnosis is the process of identifying the specific disease that is causing a person's symptoms. The diagnosis can be particularly problematic due to the presence of non-specific symptoms and indications. The accurate diagnosis of diseases is crucial for effective therapy. ML is a field that utilises past training data to make predictions about disease diagnosis. Scientists have developed multiple ML algorithms to effectively diagnose different diseases. ML enables machines to acquire knowledge and skills without the need for explicit programming. Utilising ML methods, a model can be developed to accurately forecast the early detection of diseases and offer potential remedies. Early detection and efficient treatment are the most effective means of reducing mortality rates caused by any illness. Consequently, the majority of medical experts are attracted to emerging predictive model technologies that utilise ML algorithms for disease prediction.

ML models acquire knowledge from patterns in the training samples through unsupervised learning and subsequently employ inference to provide valuable predictions. Classification techniques are widely used in the medical field to enhance the accuracy of disease identification and prediction. Conditions such as “liver cancer, chronic renal disease, breast cancer, diabetes, and cardiac syndrome" exert a substantial influence on an individual's well-being and might result in mortality if disregarded. The healthcare business will make successful decisions by uncovering concealed patterns and correlations inside the database. This study performed a comprehensive analysis to investigate the most recent advancements in ML and DL for predicting healthcare outcomes. The presentation centred around healthcare forecasting and the applicability and reliability of ML and DL techniques. Fig. 1 illustrates the review of a total of 15 papers.

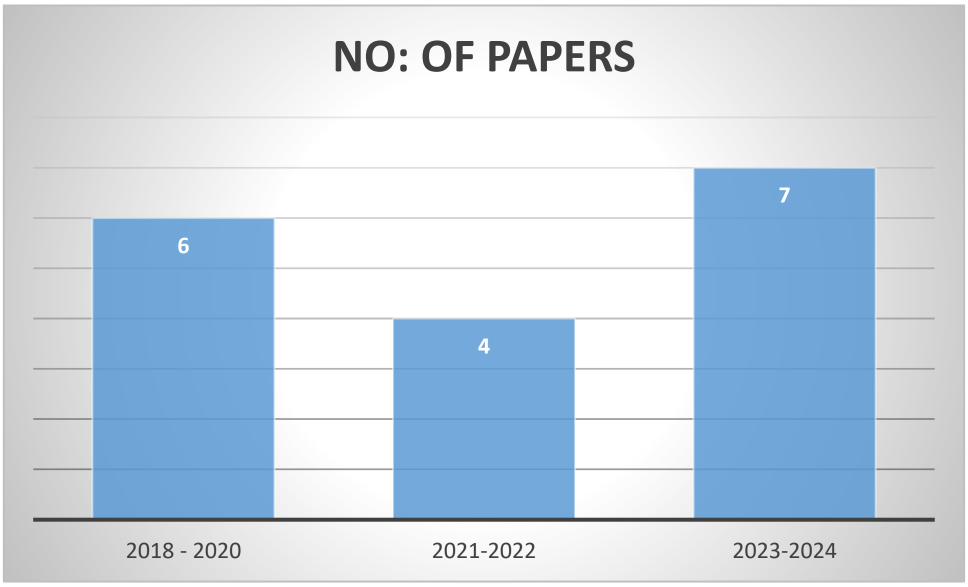

Figure 1: No: of Papers reviewed

Advancements in ML and artificial intelligence have led to the development of several classifiers and clustering algorithms, such as "K-nearest, Decision Tree, Random Forest, Support Vector Machine (SVM), Naïve Bayes, and others". These algorithms can provide a solution to the given problem. The benefits and drawbacks of each algorithm are succinctly outlined in Table I.

Table 1: Pros and Cons of Algorithms

ALGORITHMS	PROS	CONS
SVM	Data that is linear and nonlinear can be managed by it. Overfitting is less likely to occur than it was before. When dealing with high-dimensional data, scale up.	In the presence of a large dataset, performance will decrease. Selecting an appropriate kernel function can be a challenging task. When a dataset has noise, they do not perform well.
Naïve Bayes	It is simple for extremely large datasets. Capable of managing both discrete and continuous data. Both binary and multi-classification can be accomplished using this tool. A lack of sensitivity to superfluous characteristics	Particularly computationally intensive, particularly for models that contain a large number of variables Due to the fact that Naïve Bayes models are overly simplistic, there are instances in which models that have been adequately trained and optimised outperform them.
KNN	Models are inexpensive and simple to put into action. Utilised for each of the classification and regression processes.	Computation is quite intensive. Classifying unknown records is comparatively costly. It is highly subtle to inappropriate factors.

	Performs faultlessly when applied to problems involving many classes	
K-Mean	The simplicity of execution Hierarchical clustering is less effective than other methods when dealing with huge variables.	Estimating the K value is challenging. Cluster performance declines when the clusters have a spherical shape. Prone to being influenced by extreme values and random fluctuations.
Decision Tree	Applicable for both regression & classification tasks. Effortlessness in manipulating both numerical and category data	Possibly, overfitting arises when the tree is repeatedly constructed. Interpreting larger trees becomes challenging.
Random Forest	This tool is applicable for both regression and classification tasks. Address the issue of overfitting in the decision tree.	Require a significant amount of time for training due to its intricacy.
Logistic regression	Computational efficiency Ease of regularization For input features, no scaling is required	It is challenging to find a solution for an issue that is not linear.
Deep Learning	Detects automatically features Can be applied on different data types.	Require GPUs for the purpose of training. Training is expensive due to the intricate nature of the data models.

Despite the remarkable progress made in recent years, both ML and) DL still have significant challenges to overcome in order to successfully address the underlying issues afflicting healthcare systems. This study of Botlagunta et al., (2023) [12] discussed the issues related to adopting ML and DL methodologies in healthcare prediction.

The fundamental difficulty that must be addressed is the management of the Biomedical Data Stream. A substantial volume of novel medical data is being generated at a quick pace, and the healthcare sector is undergoing rapid evolution. Real-time biological signals that can be measured include blood pressure, oxygen saturation, and glucose levels [13]. Although several iterations of DL architecture have made efforts to tackle this issue, numerous obstacles still exist before efficient analyses of swiftly changing, enormous quantities of streaming data can be carried out. These issues encompass concerns related to memory usage, selecting relevant features, handling missing data, and managing computational complexity. ML and DL face the difficulty of dealing with the intricate nature of the healthcare field [14].

The complexities faced in healthcare and biomedical research surpass those encountered in other domains. There is still a significant amount of information that remains unknown on the origins, transmission, and treatments for numerous very varied diseases. Acquiring adequate data can be challenging due to the scarcity of patients at times. A resolution to this problem can potentially be discovered [15]. Due to the limited number of patients, it is necessary to do thorough patient profiling, employ advanced data processing techniques, and integrate other datasets [16]. Researchers can independently process each dataset using the suitable deep learning technique and subsequently consolidate the outcomes into a single model to extract patient data.

CONCLUSION:

The analysed studies have demonstrated that artificial intelligence techniques, specifically ML and DL, have a substantial impact on effectively detecting diseases and aiding in the prediction and analysis of healthcare data.

This is achieved by connecting several clinical records and reconstructing a patient's medical history using this information. This work enhances research in the domain of healthcare predictive analytics by employing ML and DL methodologies. It also serves as a valuable reference for other scholars and researchers, so contributing to the existing literature and facilitating future studies in this subject.

REFERENCES:

[1]. Bakator, M., & Radosav, D. (2018). Deep learning and medical diagnosis: A review of literature. *Multimodal Technologies and Interaction*, *2*(3), 47.

[2]. MacEachern, S. J., & Forkert, N. D. (2021). Machine learning for precision medicine. *Genome*, *64*(4), 416-425.

[3]. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, *7*, 81542-81554.

[4]. Moradi, H., Al-Hourani, A., Concilia, G., Khoshmanesh, F., Nezami, F. R., Needham, S., ... & Khoshmanesh, K. (2023). Recent developments in modeling, imaging, and monitoring of cardiovascular diseases using machine learning. *Biophysical Reviews*, *15*(1), 19-33.

[5]. Myszczynska, M. A., Ojamies, P. N., Lacoste, A. M., Neil, D., Saffari, A., Mead, R., ... & Ferraiuolo, L. (2020). Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nature Reviews Neurology*, *16*(8), 440-456.

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

[7]. Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y. R., & Suraj, R. S. (2021, January). Heart disease prediction using hybrid machine learning model. In *2021 6th international conference on inventive computation technologies (ICICT)* (pp. 1329-1333). IEEE.

[8]. Ramesh, T. R., Lilhore, U. K., Poongodi, M., Simaiya, S., Kaur, A., & Hamdi, M. (2022). Predictive analysis of heart diseases with machine learning approaches. *Malaysian Journal of Computer Science*, 132-148.

[9]. Arumugam, K., Naved, M., Shinde, P. P., Leiva-Chauca, O., Huaman-Osorio, A., & Gonzales-Yanac, T. (2023). Multiple disease prediction using Machine learning algorithms. *Materials Today: Proceedings*, *80*, 3682-3685.

[10]. Alqahtani, T. M. (2023). Big Data Analytics with Optimal Deep Learning Model for Medical Image Classification. *Comput. Syst. Sci. Eng.*, *44*(2), 1433-1449.

[11]. Caruccio, L., Cirillo, S., Polese, G., Solimando, G., Sundaramurthy, S., & Tortora, G. (2024). Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot. *Expert Systems with Applications*, *235*, 121186.

[12]. Botlagunta, M., Botlagunta, M. D., Myneni, M. B., Lakshmi, D., Nayyar, A., Gullapalli, J. S., & Shah, M. A. (2023). Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms. *Scientific Reports*, *13*(1), 485.

[13]. Ngiam, K. Y., & Khor, W. (2019). Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, *20*(5), e262-e273.

[14]. Sawhney, R., Malik, A., Sharma, S., & Narayan, V. (2023). A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease. *Decision Analytics Journal*, *6*, 100169.

[15]. Srivastava, D., Pandey, H., & Agarwal, A. K. (2023). Complex predictive analysis for health care: a comprehensive review. *Bulletin of Electrical Engineering and Informatics*, *12*(1), 521-531.

[16]. Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making*, *19*(1), 1-16.