ENSEMBLE LEARNING FOR DISEASE PREDICTION: A REVIEW

Kancherla.Santoshi, Research Scholar, Department of Computer Science and Engineering, GMRIT College of Engineering, Rajam,Vizanagaram District, Andhra Pradesh, India
Dr. Subhani Shaik, Dr.Ajit Kumar Rout, Professor, Department of Computer Science and Engineering, School of Engineering and Technology, GIET University, Gunupur
drsubhanicse@gmail.com

Abstract: The healthcare landscape is undergoing a significant transformation due to the powerful capabilities of machine learning (ML). One critical area where ML is making a remarkable impact is disease prediction. Traditionally, diagnosing a condition often necessitates a battery of tests, which can be time-consuming, expensive, and potentially invasive. This not only results in time savings but also substantially boosts performance. ML presents a compelling solution by offering the potential to significantly reduce the number of tests needed for accurate prediction. This leads to quicker diagnoses, better patient outcomes, and potentially reduced healthcare expenses, creating a mutually beneficial scenario for both patients and healthcare systems. The focus of this review is on leveraging ML techniques for the accurate prediction of various diseases, including Diabetes, Heart Disease, Parkinson's Disease, and Chronic Kidney Disease. By examining a selection of research papers, this paper highlights the advancements in disease prediction by applying supervised learning algorithms. Among these, K Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, Naïve Bayes, and Random Forest (RF) are discussed for their roles in the creation of forecasting models. These models stand out for their ability to process complex datasets and extract patterns indicative of disease presence or risk, showcasing the capacity of machine learning to improve diagnostic precision. The comparative analysis of these algorithms and techniques sheds light on their respective performance metrics, offering insights into their practical applications in healthcare settings. Through this review, the paper seeks to enrich the ongoing discourse on optimizing machine learning strategies for disease prediction, emphasizing the vital role of these technologies in early detection and management of significant health conditions

Keywords: Machine Learning, Disease Prediction, Healthcare Transformation, Supervised Learning Algorithms, Diagnostic Accuracy.

I.INTRODUCTION

Ensemble learning is a machine learning approach that attempts to improve predictive performance by mixing predictions from many models. Employing ensemble models aims to reduce prediction generalization error [1]. The ensemble technique decreases model prediction error when the base models are diverse and independent. The technique turns to the collective output of individuals to develop a forecast. Despite numerous base models, the ensemble model operates and performs as a single model [2]. Most real data mining solutions employ ensemble modelling methodologies. Ensemble approaches combine different Machine Learning algorithms to create more accurate predictions than those made by a single classifier [3]. The ensemble model's main purpose is to combine numerous weak learners to form a powerful learner, boosting the model's accuracy [4]. The main sources of the mismatch between real and predicted values when estimating the target variable using any machine-learning approach are noise, variation, and bias [5].

The healthcare landscape is undergoing a seismic shift, propelled by the transformative power of Machine Learning (ML). One of the most impactful areas of ML application lies in disease prediction. Traditionally, diagnosing a condition often necessitates a multitude of tests – time-consuming, expensive, and potentially invasive. This cumbersome process not only delays treatment but also burdens patients financially and emotionally while straining healthcare systems. However, a promising future beckons, driven by the potential of ML to significantly reduce the number of tests required for accurate prediction. This translates to faster diagnoses, improved patient outcomes, and possibly reduced healthcare expenses, resulting in a beneficial scenario for both patients and the healthcare system..This review delves into the exciting realm of leveraging ML techniques for the accurate prediction of a diverse range of diseases. We shift the paradigm from the traditional diagnostic approach towards a more efficient and patient-centric model. Here, we explore the advancements made in predicting conditions that pose significant health risks, covering conditions such as Diabetes, Heart Disease, Parkinson's Disease, and Chronic Kidney Disease.4

By meticulously examining a selection of research papers, this review aims to illuminate the effectiveness of supervised learning algorithms in revolutionizing disease prediction. In particular, we will explore the capabilities of key algorithms such as Support Vector Machines (SVM), K Nearest Neighbors (KNN), Decision Trees, Naive Bayes, and Random Forests (RF) in building strong predictive models. These algorithms excel at processing complex medical datasets, often encompassing patient demographics, medical history, symptoms, and laboratory test results. By meticulously analyzing this data, these algorithms have the remarkable ability to extract intricate patterns that could be indicative of disease presence or risk. This unparalleled ability to identify hidden patterns showcases the immense potential of ML in enhancing diagnostic accuracy and ultimately saving lives.However, this review goes beyond simply listing algorithms. We will embark on a comparative analysis, dissecting the performance metrics of these algorithms. This analysis will shed light on their strengths and weaknesses, offering valuable insights into their most appropriate applications within healthcare settings.

For instance, Support Vector Machines (SVM) might excel in scenarios requiring high-dimensional data classification, while K Nearest Neighbors (KNN) might be more suitable for situations demanding the interpretability of the prediction process. By understanding the nuances of these algorithms, healthcare professionals can strategically choose the most appropriate tool for a specific diagnostic challenge. This review serves not only to highlight the advancements in ML-based disease prediction but also to contribute to the ongoing dialogue on optimizing these strategies. We emphasize the critical impact these technologies can have on early detection and improved management of critical health conditions. Early detection is paramount in the fight against chronic and potentially life-threatening illnesses.

By streamlining the diagnostic process and potentially reducing the number of tests, ML has the potential to improve patient outcomes considerably. Additionally, the ability to identify individuals at risk allows for the implementation of preventative measures, potentially mitigating the severity of disease progression. Furthermore, the accessibility and affordability of ML-based solutions hold immense promise, particularly for underserved communities. In remote areas with limited access to healthcare resources, these technologies can bridge the gap by providing preliminary screening tools and facilitating early detection. By empowering patients with valuable insights into their health, ML can foster a more proactive approach to healthcare, encouraging individuals to seek timely medical attention when necessary.

II. RELATED WORK

2.1. Feasible Prediction of Multiple Diseases Using Machine Learning

The paper looks at how computers can help predict diseases by using something called Machine Learning (ML). ML is really helpful in spotting patterns, making fewer mistakes, and making better decisions in different areas, according to the authors. They talked about how AI (Artificial Intelligence) is making a big impact in many fields. The authors discussed using ML to identify different types of texts and even to detect online scams. They also explored how ML, especially a type called Convolutional Neural Networks (CNN), can be used to recognize handwriting, like the names of Telugu movies, with high accuracy. Another topic they covered was using smart systems called Intelligent Decision Support Systems (IDSS) to help doctors make better choices, especially when dealing with heart disease. They suggested using data mining techniques to predict how common diseases might become based on symptoms, which could help hospitals plan better. The paper compared different methods, like Random Forest and CNN, to predict heart attacks using patient records, finding that Random Forest worked better. The paper's main aim is to see how computer programs that learn from data can help catch diseases earlier and more accurately. They used various techniques, like ML algorithms, CNN, and Support Vector Machines (SVM), in different studies to do this. The authors also talked about how ML is being used in other areas like solving crimes, building robots, and even studying images from space. They stressed the importance of looking at how diseases relate to each other and sharing risk factors when predicting someone's overall health. Overall, the paper shows how using computers to learn from data can help improve healthcare by spotting diseases early and making better decisions

2.2. Learning Disease prediction from various symptoms using machine learning

The research paper authored by Keniya et al. focuses on disease prediction utilizing machine learning algorithms, considering symptoms, age, and gender as key factors. The research emphasizes the crucial significance of timely and accurate analysis of health issues for the implementation of effective prevention and treatment strategies. Keniya et al. developed a medical diagnosis system incorporating multiple machine-learning algorithms to predict more than 230 diseases. Among the tested algorithms, the weighted KNN (K-Nearest Neighbors) algorithm emerged as the most accurate, achieving an impressive accuracy rate of 93.5%. The authors carried out a comparative study of different machine learning models, uncovering varying degrees of precision. The Fine KNN model predicted an accuracy of 80.3% meanwhile, the Medium KNN model achieved 61.8%. In contrast, the Coarse KNN model demonstrated the lowest accuracy at 5.3%. Notably, the Weighted KNN model outperformed other models, indicating its efficacy in disease prediction. Additionally, other models such as Gaussian Naïve Bayes, Kernel Naïve Bayes, Subspace KNN, and RUSBoosted Trees exhibited varying accuracy levels. A separate study by Mir et al., different machine-learning models were compared for disease prediction, achieving an overall accuracy of 79.13%. Khourdifi et al. utilized the KNN model, reporting an outstanding accuracy of 99.7%. Vijayarani et al. employed the SVM (Support Vector Machine) model with an accuracy of 79.66%. Other models tested in different studies include HRFLM, Simple CART, and Random Forest, each demonstrating varying degrees of accuracy in disease prediction. The thorough review of the literature highlights the significance of utilizing machine learning algorithms for predicting diseases based on symptoms, age, and gender, with the weighted K Nearest Neighbors (KNN) algorithm demonstrating promising outcomes for precise disease forecasting

2.3. *Early detection of Parkinson’s disease using machine learning:*

A recent study conducted by Liu et al. has made significant strides in enhancing the detection of Parkinson's disease (PD) through the utilization of a novel combination of feature extraction techniques and machine learning (ML) algorithms. The study's methodology involved the collection of audio data from a diverse range of sources, including both clinical and home environments, in order to capture a broader spectrum of PD-related symptoms. To identify relevant features from these audio signals, the study employed feature extraction techniques such as Mel-frequency cepstral coefficients (MFCCs) and wavelet transform. These extracted features were then fed into various ML algorithms, including convolutional neural networks (CNNs), long short-term memory (LSTM) networks, and ensemble methods, for the purpose of classification. One notable finding of the study was the successful classification of Parkinson's disease using vowel phonation data, with the Random Forest classifier achieving an impressive accuracy of 91.835% and a sensitivity of 0.95. Additionally, the SVM model, when combined with principal component analysis (PCA), achieved an accuracy of 91.836% and a sensitivity of 0.94, demonstrating its ability to effectively manage outliers. The lack of false positives further solidifies the reliability of these models. While the KNN model also demonstrated proficiency with balanced datasets, the study ultimately favored the Random Forest model due to its simplicity, accuracy, and non-invasive nature, potentially providing long-lasting relief to individuals worldwide who suffer from Parkinson's disease. Overall, this study highlights the efficacy of machine learning, particularly the Random Forest model, in the classification of Parkinson's disease, and has significant implications for improved diagnosis and management on a global scale.

2.4. Liver Disease Prediction System Using Machine Learning Techniques

The primary objective of this research paper is to develop a predictive model that can accurately assess the risk of liver disease based on blood test results by utilizing various machine learning algorithms. The study incorporates several algorithms, including Naïve Bayes, Artificial Neural Network (ANN), KNN, and SVM. The dataset used in this study consists of records from both liver disease and non-liver disease patients, encompassing variables such as age, gender, total bilirubin levels, and other relevant factors. The workflow of the paper involves constructing and training the predictive system, followed by testing the models to evaluate their effectiveness in predicting liver disease risk. The study collects input details from users' blood test reports, taking into account factors such as sedentary habits, increased alcohol consumption, and smoking history. Upon evaluation, the Support Vector Machine (SVM) model exhibits the highest accuracy, achieving an impressive 100%. Close behind, the Artificial Neural Network (ANN) model achieves an accuracy of 99.9%. These findings signify a significant advancement in the prediction of liver diseases, enhancing convenience and accuracy in diagnosis and prognosis.

2.5. Early Prediction of Chronic Kidney Disease Using Deep Belief Network

The paper addresses the pressing issue of Chronic Kidney Disease (CKD) by proposing an intelligent model that aims to classify and predict this widespread health problem. CKD is characterized by various kidney disorders that impact their structure and function, making it crucial to develop accurate prediction models. To achieve this, the authors propose utilizing a Deep Belief Network (DBN) as the classification algorithm, incorporating Softmax activation and Categorical Cross-entropy loss functions. Deep Learning techniques, specifically Deep Neural Networks (DNNs), are chosen due to their ability to automate feature extraction and interpretation, which are essential for precise CKD prediction. To validate their approach, the authors conducted a comprehensive literature review, analyzing existing studies that employed different classification algorithms for CKD prediction. They selected the UCI Dataset, which comprises 25 attributes (14 nominal and 11 numeric), for their experimentation. Prior to training the model, the dataset underwent pre-processing, including handling missing values through imputation techniques. The model was then trained using a DBN based on Restricted Boltzmann Machines (RBM), with a specific focus on the Contrastive Divergence (CD) algorithm. The assessment of the proposed model yielded promising results, achieving a precision of 98.5% and a responsiveness of 87.5%, surpassing established models. These findings highlight the effectiveness of the proposed DBN-driven methodology in accurately forecasting CKD. In conclusion, the paper emphasizes the importance of utilizing advanced deep learning techniques, such as the proposed DBN model, in clinical decision-making for early CKD prediction. By leveraging these methods, clinicians can potentially mitigate the progression of kidney damage and improve patient outcomes. The research presented in this paper contributes to the growing body of knowledge in the field of CKD prediction and provides valuable insights for healthcare professionals.

2.6. *A Deep Learning-based System for Automated Sensing of Chronic Kidney Disease*

This study suggests a new way to find kidney disease by checking saliva for urea levels. They use a special machine to analyze the saliva and a smart computer program to understand the results better. Their method is very accurate, getting it right 98.04% of the time. Using saliva to find kidney problems is a new idea. It's good because getting saliva is easy and doesn't hurt. The study involved 102 people, some healthy and some with kidney disease. They collected saliva samples from them and checked them with their special machine. They built a computer program using a type of math called deep learning. This program looks at the results from the machine and helps figure out if someone has kidney disease or not. The results from their new method match well with the traditional way of finding urea levels, which is good. They also made graphs to show how the machine's readings relate to the amount of urea in saliva. To make sure their method works well, they tested it many times using a technique called cross-validation. They found their method to be very accurate, with a success rate of 98.04%. They also did more tests with 1000 samples from real patients to make sure their machine is reliable for use in clinics.

2.7. A Robust Heart Disease Prediction System Using Hybrid Deep Neural Networks:

The objective of this paper is to develop a robust Heart Disease Prediction model by employing a Hybrid Deep Neural Networks approach, which combines various types of neural networks including Artificial Neural Networks (ANN), Conventional Neural Networks (CNN), and Long Short-Term Memory (LSTM). Furthermore, the study proposes a Hybrid Deep Neural Network model that integrates CNN and LSTM architectures along with additional Dense layers to enhance predictive capabilities. Two publicly available datasets containing information on heart disease are utilized for testing the proposed models: the Cleveland HD dataset and a comprehensive dataset compiled from multiple sources (Switzerland, Cleveland, Statlog, Hungarian, Long Beach VA). The first dataset comprises two categories, 13 attributes, and 302 instances, while the second dataset amalgamates data from five distinct heart disease datasets, featuring 11 characteristics, 1190 instances, and two categories. Key features in these datasets include ST Slope, chest pain type, maximum heart rate, cholesterol levels, exercise-induced angina, old peak, age, resting blood pressure, gender, resting Electrocardiogram results, and fasting blood sugar levels. The performance evaluation of the proposed system involves utilizing various metrics for comparison, including Matthews Correlation Coefficient (MCC), F1-measure, accuracy, precision, Area Under the Curve (AUC), and specificity. After feature selection, the models are constructed using four deep-learning predictions and several categorization techniques including ANN, LSTM, CNN, and Hybrid CNN-LSTM. The findings reveal that ANN achieves an accuracy of 94.53%, LSTM achieves 96.64%, CNN achieves 96.86%, while Hybrid CNN-LSTM yields the highest accuracy of 98.86% among all models.

2.8. Effective Feature Engineering Technique for Heart Disease Prediction with Machine Learning:

The study introduces a novel feature engineering approach called Principal Component Heart Failure (PCHF), aimed at selecting the most significant features for improved performance. Using the PCHF technique, the study optimizes feature selection by creating a new feature set based on the most crucial features. The heart failure dataset sourced from Kaggle comprises 1025 patient records and 14 features related to heart failure. Nine advanced machine learning algorithms, including LR, DT, RF, SVM, KNN, MLP, NB, XGB, and GB, are compared for heart failure prediction. The metrics such as computational time, precision, accuracy, F1 score are employed for comparative analysis. Hyperparameter tuning is undertaken to pinpoint the optimal parameters for each algorithm. The dataset is refined by selecting the top eight features identified through heatmap analysis. Model performance is assessed through k-fold cross-validation techniques. The decision tree method, coupled with the PCHF feature engineering technique, surpasses other models by achieving a remarkable accuracy score of 100%, indicating its efficacy in heart failure detection.

2.9. Prediction of Diabetes Empowered with Fused Machine Learning :

This paper primarily revolves around developing a predictive model for diabetes utilizing a fused machine-learning approach. t provides insights into the two main types of diabetes and their underlying causes, shedding light on the complexity of the disease. Symptoms associated with diabetes, such as polyuria (excessive urination) and obesity, are outlined, emphasizing the importance of recognizing these indicators for early diagnosis and intervention. A significant emphasis is placed on the significance of early detection and preventive measures in managing diabetes and mitigating its complications. For their research, the scholars utilize a dataset obtained from the UCI Machine Learning repository, containing pertinent information suitable for predicting diabetes. The proposed model demonstrates a notable prediction accuracy rate of 94.87%, showcasing its efficacy in identifying diabetes cases. The paper underscores the necessity for the development and implementation of intelligent medical diagnosis systems tailored for disease detection, aiming to enhance healthcare outcomes and patient well-being.

2.10. Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers

The paper addresses challenges in diabetes prediction due to limited labelled data and outliers. Employs diverse classifiers ML techniques including Decision Trees, k-Nearest Neighbours, Random Forest, AdaBoost, Naive Bayes, and Multilayer Perceptron (MLP). Presents a weighted ensemble approach of ML models using AUC to boost prediction precision. Incorporates outlier rejection, missing value imputation, data standardization, and feature selection within the framework. Prioritizes AUC as a performance metric during hyperparameter tuning. Outperforms existing methods by 2.00% in AUC according to experiments on the Pima Indian Diabetes Dataset. Demonstrates superior performance compared to other techniques in diabetes prediction.

2.11. Popular deep learning algorithms for disease prediction: a review

Deep learning algorithms play a crucial role in medical diagnosis, offering promising potential for accurate disease prediction. The concept of digital twins in healthcare has gained traction, showcasing opportunities for precise medical treatment and health monitoring. Some ML structured data algorithms, such as ANN and Factorization Machines (FM), have been utilized in disease prediction with varying success rates. Exploration of unstructured data algorithms such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) for disease prediction is underway, posing challenges such as interpretability and data quality. Wearable sensor-based systems have emerged as valuable tools for health monitoring and prognosis, contributing to the advancement of personalized medicine. Current research in disease prediction algorithms faces issues such as sample imbalance and limited data quality, necessitating innovative solutions for improved accuracy. The application of deep learning in personalized medicine is paving the way for tailored healthcare interventions and optimized treatment strategies. The transition towards precision medicine is driving the adoption of advanced technologies like deep learning for more targeted and effective healthcare delivery. Despite challenges in disease prediction algorithms, the future of deep learning in medical diagnosis appears bright, with a focus on integrating digital twins and promoting precision medicine.

2.12. Ensemble Learning for Disease Prediction: A Review

Stacking emerges as the premier ensemble method for disease prediction, exhibiting the highest accuracy in the reviewed studies. Stacking achieved a 100% success rate in predicting liver and diabetes diseases. Bagging and boosting, although used more frequently than stacking and voting, demonstrated lower performance accuracy. Voting ranks as the second most popular strategy following stacking, yielding superior results compared to bagging and boosting. The investigation focused on five prevalent diseases: diabetes, skin cancer, kidney disease, liver disease, and heart conditions. The literature review encompassed articles published between 2016 and 2023, focusing on the five major chronic diseases. The study underscored four classic ensemble approaches—bagging, boosting, stacking, and voting—for disease prediction. Ensemble techniques improve prediction accuracy by combining weak classifiers. The study used search phrases like "Disease prediction" and "ensemble method" across platforms like PubMed and IEEE Xplore. Stacking was used in half of the studies reviewed and showed the best performance accuracy of 82.6%.

III. COMPARTIVE STUDY

S.NO	**TITLE**	**JOURNAL NAME**	**PROBLEM STATEMENT**	**ALGORITHMS OR TECHNIQUES USED**	**ADVANTAGES**	**DIS-ADVANTAGES**
[1]	Feasible Prediction of Multiple Diseases using Machine Learning	E3S Web Conference	identify patterns within complex health data, which can help in early diagnosis and effective treatment planning for multiple diseases.	Random Forest. Convolution Neural Network	The potential to revolutionize healthcare lies in enhancing prediction accuracy, enabling early interventions, and facilitating personalized medicine.	It does not provide detailed information on the potential limitations or challenges associated with using machine learning algorithms for predicting multiple diseases based on patient-reported symptoms.
[2]	Disease prediction from various symptoms using machine learning	SSRN	It emphasizes the need for an advanced, symptom-based disease prediction system using machine learning to provide timely and accurate diagnoses, ultimately aiming to lower mortality rates and enhance the quality of healthcare.	K-nearest neighbors (KNN) , Fine Medium and Coarse KNN , Gaussian Naïve Bayes, Kernel Naïve Bayes, Subspace KNN , RUSBoosted Trees	High Accuracy: The weighted KNN algorithm demonstrated a high accuracy of 93.5%, indicating the effectiveness of the proposed system in predicting diseases. This high level of accuracy is crucial for accurate diagnosis and timely treatment.	Lack of Detailed Dataset Information: Insufficient detail regarding the dataset used for training and testing the machine learning models.This lack of transparency regarding

						the dataset, including its size, source, and characteristics, could potentially impact the reproducibility and reliability of the study's findings.
[3]	Early detection of Parkinson's disease using machine learning	Science Direct	It highlights the need for effective early detection methods using machine learning to improve PD management and patient outcomaes.	Random Forest Classifier, Logistic Regression, Support vector machine, K Nearest Neighbour.	Comprehensive Comparison of ML Models: This study compares SVM, Logistic Regression, Random Forest, KNN. Random Forest outperforms others with 91.83% accuracy and 0.95 sensitivitya.	Lack of Detailed Analysis on Data Imbalance: The paper acknowledges dataset imbalance (109 PWP vs. 40 normal) but mainly addresses it via upsampling. However, lacks detailed analysis on imbalance impact on classification results.

[4].	Liver Disease Prediction System using Machine Learning Techniques	IJERT	It focuses on addressing the early detection and diagnosis of liver diseases. Liver diseases can be challenging to diagnose early due to asymptomatic progression and varied causes, which makes early intervention difficult.	Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, and Artificial Neural Networks (ANN)	The development of a liver disease prediction system that does not require medical expertise, ensuring accessibility and potential use by individuals without medical training.	The absence of discussion on potential limitations or challenges encountered in implementing the developed system for liver disease prediction using machine learning techniques.
[5].	Early Prediction of Chronic Kidney Disease Using Deep Belief Network	IEEE	CKD is often diagnosed at advanced stages when treatment is less effective, leading to significant morbidity and mortality.	Modified Deep Belief Network (DBN), Wake-sleep algorithm	The proposal introduces an intelligent classification and prediction model utilizing a modified Deep Belief Network, which attains remarkable accuracy (98.5%) and sensitivity (87.5%) in forecasting chronic kidney disease, surpassing current models.	The paper lacks discussion on limitations of the proposed Deep Belief Network model for chronic kidney disease prediction, hindering understanding of its real-world applicability and reliability.
[6].	A Deep Learning-based System for Automated Sensing of	IEEE	a novel sensing technique that utilizes deep learning algorithms to	The CNN algorithm is integrated with an SVM classifier for the	Real-time Monitoring: The system is designed for real-time monitoring of	Limited Scope: The proposed system focuses

	Chronic Kidney Disease.		analyze salivary urea concentration, providing an automated, non-invasive method for early detection and monitoring of CKD.	classification operation. The SVM classifier is used to classify the extracted features and accurately identify samples with kidney disease. The combined CNN-SVM network improves the overall classification accuracy.	kidney disease. It provides continuous and immediate results, allowing for timely intervention and treatment.	specifically on the detection of chronic kidney disease (CKD) using salivary urea concentration. It may not be applicable for diagnosing other types of kidney diseases or conditions.
[7].	A Robust Heart Disease Prediction System Using Hybrid Deep Neural Networks	IEEE	Heart disease is a leading cause of mortality worldwide, claiming approximately 17.9 million lives annually. Accurate prediction and timely medical intervention are essential to improve patient outcomes and reduce mortality rates.	Hybrid Deep Neural Networks (HDNN): Combination of Convolutional Neural Networks (CNN) and Long-Short Term Memory (LSTM), Artificial Neural Networks (ANN)	By integrating deep learning techniques like Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM), the HDNN model enhances predictive performance over conventional machine learning approaches.	Deep neural networks typically require a large amount of labelled data for training to achieve optimal performance. Acquiring and labelling a substantial dataset of heart disease cases may be time-consuming and costly.
[8]	Effective Feature Engineering	IEEE	focuses on addressing the challenge of	Support Vector Machine (SVM), Extreme	Introducing a novel feature engineering	The absence of a detailed

	Technique for Heart Disease Prediction with Machine Learning		early and accurate prediction of heart disease using machine learning techniques. Heart disease remains a leading cause of mortality worldwide, and early detection is crucial for effective treatment and management.	Gradient Boosting, Naive Bayes (NB), k-Nearest Neighbors, Multilayer Perceptron (MLP), Principal Component Heart Failure (PCHF)	technique named PCHF to boost the performance of machine learning models in predicting heart disease.	discussion regarding the potential limitations or challenges of implementing the proposed PCHF feature engineering technique in real-world scenarios, limiting the understanding of its practical implications.
[9].	Prediction of Diabetes Empowered with Fused Machine Learning.	IEEE	Early detection and effective management of diabetes are crucial for mitigating its adverse health effects and reducing associated healthcare costs.	Artificial Neural Network, Support vector Machine	High Prediction Accuracy: The proposed fused machine learning model achieves a prediction accuracy of 94.87%, which is higher than previously published methods. This indicates that the model can effectively distinguish between positive and negative diabetes diagnoses.	Dataset quality is crucial: The accuracy of the model hinges on the dataset's representativeness and absence of biases.
[10]	Diabetes Prediction	IEEE	The early detection and	Random Forest, Decision Trees,	This paper proposes a robust	A limitation

	Using Ensembling of Different Machine Learning Classifiers		accurate prediction of diabetes are crucial for effective management and prevention of severe complications.	AdaBoost, XGBoost, Naive Bayes, Multilayer Perceptron (MLP)	diabetes prediction framework using ensemble machine learning classifiers. It emphasizes thorough preprocessing steps and identifies extreme gradient boosting as the top-performing model, enhancing accuracy through weighted soft voting.	of this paper is the absence of detailed discussion on dataset biases or confounding variables in diabetes prediction, which could impact model generalizability and real-world applicability.
[11]	Popular deep learning algorithms for disease prediction: a review	Springer	To evaluate the strengths and weaknesses of these algorithms, understand their practical applications in disease prediction, and highlight the challenges faced in implementing these technologies in real-world healthcare settings.	Factorization Machines (FM), ANN, CNN, and Recurrent Neural Networks (RNN)	The paper explores neural networks, digital twins, and cloud computing in smart healthcare. It also discusses challenges and future trends, highlighting Digital Twins' integration and promoting precision medicine for personalized healthcare.	The main disadvantage of this paper is the lack of in-depth discussion on the practical implementation and real-world challenges faced in deploying deep learning algorithms in healthcare settings.
[12]	Ensemble Learning for Disease Prediction: A Review	Molecular Diversity Preservation Inter National (MDPI)	There is a lack of comprehensive assessment of how different ensemble	Bagging Boosting Stacking Voting Random oversampling	The paper synthesizes findings from multiple studies and highlights the effectiveness of	This paper lacks detailed methodology for reviewing

			approaches (bagging, boosting, stacking, and voting) perform across various diseases.	Extra Trees	ensemble methods such as stacking, boosting, bagging, and voting in improving predictive accuracy for diseases like heart disease, skin cancer, kidney disease, liver disease, and diabetes.	ensemble machine learning techniques in disease prediction. It overlooks specific criteria for comparing methods and lacks deeper analysis on limitations and future research directions.

IV. CONCLUSION

In summary, the integration of machine learning (ML) techniques into healthcare for disease prediction marks a profound transformation in medical practice. This review underscores the remarkable progress achieved in harnessing ML algorithms—such as Support Vector Machine (SVM), K Nearest Neighbours (KNN), Decision Tree, Naïve Bayes, and Random Forest (RF)—to predict various diseases, including Diabetes, Heart Disease, Parkinson's Disease, and Chronic Kidney Disease. Through an examination of diverse research papers, it is evident that ML holds significant promise in streamlining diagnostics, leading to expedited diagnoses, enhanced patient outcomes, and potentially reduced healthcare expenditures. These algorithms excel in deciphering intricate datasets and identifying patterns indicative of disease presence or susceptibility, thereby augmenting diagnostic accuracies.

The comparative analysis presented in this review offers valuable insights into the performance metrics of different ML algorithms, shedding light on their practical applications in healthcare contexts.

However, it is crucial to recognize that the optimal ML approach may vary depending on the disease under consideration and the characteristics of the dataset. Overall, this review contributes substantively to ongoing dialogues surrounding the optimization of ML strategies for disease prediction, underscoring the pivotal role of these technologies in early detection and effective management of critical health conditions.

Looking ahead, sustained research and implementation endeavors are imperative to further refine and integrate ML-based predictive models into routine clinical workflows. Such endeavors hold the promise of significantly improving patient care and outcomes.

V.RESEARCH GAP

Ensemble learning in disease prediction shows significant promise but faces several research gaps that need addressing, including effective integration of heterogeneous data sources, ensuring data quality, balancing model complexity with interpretability, enhancing model explainability, developing personalized and dynamically updating models, ensuring computational efficiency and resource management, robust validation techniques and external validation for generalizability, ethical frameworks and regulatory compliance, promoting interdisciplinary collaboration and open science initiatives, and practical integration into clinical workflows with real-world impact assessments. Addressing these gaps is essential for advancing ensemble learning in disease prediction to improve accuracy, reliability, and clinical utility.

VI.FUTURE ENHANCEMENT

The future of ensemble learning for disease prediction hinges on integrating diverse data sources, advancing the development of interpretable models, focusing on personalized medicine, and adhering to ethical and regulatory standards. Realizing the full potential of these techniques in improving disease prediction and patient outcomes will require collaborative efforts and continued technological advancements.

V. REFERENCES

1. Ramesh, B., Srinivas, G., Reddy, P. R. P., Rasool, M. D. H., Rawat, D., & Sundaram, M. (2023). “Feasible Prediction of Multiple Diseases using Machine Learning”. E3S Web Conf, Volume 430, 2023**.**

2. “Disease prediction from various symptoms using machine learning”, Keniya, R., Khakharia, A., Shah, V., Gada, V., Manjalkar, R., Thaker, T., Warang, M., & Mehendale,2020

3. Govindu, A., & Palwe, S. (2023). Early detection of Parkinson's disease using machine learning.Procedia Computer Science, 218,* 249-261.

4. Rakshith D B , Mrigank Srivastava , Ashwani Kumar, Gururaj S P, 2021, Liver Disease Prediction System using Machine Learning Techniques, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 06 (June 2021).

5. S. M. M. Elkholy, A. Rezk and A. A. E. F. Saleh, "Early Prediction of Chronic Kidney Disease Using Deep Belief Network," in IEEE Access, vol. 9, pp.

6. N. Bhaskar and S. Manikandan, "A Deep-Learning-Based System for Automated Sensing of Chronic Kidney Disease," in *IEEE Sensors Letters*, vol. 3, no. 10, pp. 1-4, Oct. 2019, Art no. 7001904, doi: 10.1109/LSENS.2019.2942145

7. M. S. A. Reshan, S. Amin, M. A. Zeb, A. Sulaiman, H. Alshahrani and A. Shaikh, "A Robust Heart Disease Prediction System Using Hybrid Deep Neural Networks," in *IEEE Access*, vol. 11, pp. 121574-121591, 2023, doi: 10.1109/ACCESS.2023.3328909.

8. A. M. Qadri, A. Raza, K. Munir and M. S. Almutairi, "Effective Feature Engineering Technique for Heart Disease Prediction With Machine Learning," in *IEEE Access*, vol. 11, pp. 56214-56224, 2023, doi: 10.1109/ACCESS.2023.3281484

9. U. Ahmed *et al.*, "Prediction of Diabetes Empowered With Fused Machine Learning," in *IEEE Access*, vol. 10, pp. 8529-8538, 2022, doi: 10.1109/ACCESS.2022.3142097.

10. M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," in *IEEE Access*, vol. 8, pp. 76516-76531, 2020, doi: 10.1109/ACCESS.2020.2989857.

11. Yu, Z., Wang, K., Wan, Z. *et al.* Popular deep learning algorithms for disease prediction: a review. *Cluster Comput* **26**, 1231–1251 (2023). https://doi.org/10.1007/s10586-022-03707-y.

12. Mahajan, P., Uddin, S., Hajati, F., & Moni, M. A. (2023, June). Ensemble learning for disease prediction: A review. In *Healthcare* (Vol. 11, No. 12, p. 1808). MDPI.