Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 13, Issue 02, February 2024

Enhanced Multiple Disease Detection Using Machine Learning

N.M.Ramalingeswara Rao, V.Raghu Ram, K.Mohan Naidu,
K.Chandu, R.Sampath Kumar.
Electronics and Communication Engineering.
Godavari Institute Of Engineering and Technology(Autonomous), Rajahmundry

Abstract:
Effective detection of illnesses is an influential unmet need absent from a global scale. A rise of early diagnosis tools and a successful therapy options has been seriously compromised by the complexity of the different conditions mechanisms and underlying symptoms that impacted the patient population. A subfield of artificial intelligence called machine learning (ML) helps patients, doctors, and researchers find products to some of the aforementioned issues. This review provides an justification of Machine Learning (ML) based on studies that are relevant. For the purpose of improving patient outcomes and alleviating healthcare costs, swift detection and diagnosis of medical issues like diabetes, chronic kidney disease, liver disease, and breast cancer are essential. Machine learning have grown into a viable method for screening for and detecting chronic illnesses in recent years. We want that will assist with early disease diagnosis and treatment by creating a machine learning model for various disease belonging through this project. A significant set of data of medical records, ranging from patient demographics, medical histories, symptoms, and test findings for a diagnostic will be used to train the model that has been suggested. So excellent data quality and completeness, the dataset will be meticulously selected and preprocessed. Relevant nuances that was demonstrated to affect illness risk, such as genetic markers and decisions regarding one's lifestyle, will also be were present.

Keywords: **Machine Learning, Decision tree, Adaboost , Xgboost and Catboost and ML techniques, evaluation**.

INTRODUCTION:

Afresh developed technique designation "multiple disease detection" uses machine learning algorithms and intricate data analytics to find several diseases promptly.

the process of a examination of data from multiple accounts, incorporating imaging studies, laboratory tests, and electronic medical records (EMRs), this technology seeks providing comprehensive disease diagnostics, improving with the accuracy and efficacy of medical diagnoses [1, 2]. Because the early detection might significantly enhance patient outcomes, this technology tends to be crucial chronic diseases with a value diabetes, chronic kidney disease, liver disease, and breast cancer [3, 4].

If diabetes doesn't have a early on, it could arise in catastrophic repercussions. Diabetes is a chronic condition marked by elevated blood sugar levels. Issues like kidney disease and blindness can be got away with early intervention [5]. In a similar vein, early detection of Chronic Kidney Disease can potentially halt the disease's course and avoid kidney failure [6]. As early detection can improve liver function and reduce damage to the liver, which is why liver disease, which stretches to an range of sickness that influences the liver, can be better handled [7]. When taken care of early, breast cancer, the most frequent cancer in women the world over, has a notably greater survival rate [8].

The consumption of Multiple Disease Detection technology helps health care providers treat folks to efficiently and swiftly in addition to improve a diagnostic test. This is vital when addressing prolonged medical conditions and cutting across healthcare prices [9, 10]. More sophisticated devices for diagnosing and treating multiple diseases receive on behalf of new advancements in technology, for the aging population where multifactorial diseases like chronic obstructive pulmonary disease and cancer are growing into more common [10, 11]. The increasing comprehension of the link amongst multifactorial imperfections and cancer highlights the need for more study to make connection

common lethal cycles, whose can bring about in his peers of fresh drugs that target multiple pathways [12, 13]. Many research initiatives have greatly enhanced the specialty of machine learning's advancement of healthcare, especially while it comes to Multiple Disease Detection. As a story of the promise of such techniques in frustrating biological data analysis [14], the work of G. W. Bassel et al. [15] articulated the value of rule-based machine learning in creating functional networks, particularly in biological environments like Arabidopsis. Similarly, S. M. Weiss and N. Indurkhya [16] highlighted a range of these techniques in medical diagnostics prior to went over the details of the efficacy of rule-based machine learning for functional prediction.

Additionally, a comprehensive overview to learning classifier systems was given by R. J. Urbanowicz and J. H. Moore [18], who contained a method that blends rule- based learning with evolutionary algorithms. An opportunity to cultivate ever more sophisticated diagnostic tools that can change and push in reaction in new material is illustrated by this integration.

L. N. Castro, L. N. D. Castro, and J. Timmis [19] pointed out an uprising computational intelligence approach during their concept of Artificial Immune Systems. This methodology offering a dynamic and a powerful framework for disease detection and categorization, engaging in inspiration from the biological immune system. Fuzzy logic concepts are used in a model presented by H. F. El-Sofany and I. A. T. F. Taj-Eddin
[20] to improve the accuracy and flexibility of cloud-based medical diagnostics. That approach is especially pertinent to govern the murky waters and shifting that come with allowing medical conditions.

C. Shivade et al.'s work [21] talked about many of them of methods for with electronic health records to identify patient observable cohorts. This review signals how integral it is to accurately determine patient data in order to effectively detect and treat problems. R. Zhang et al. [22] originated automatic method of uncovering New York Heart Association classification in the field of clinical notes interpretation, proving the machine-learning capacity to understand and then create placement of unstructured clinical data.

L. S. Kumar and A. Padmapriya [23], through their research on rule-based information extraction from electronic health records, emphasized the importance of structured data processing in healthcare. Lastly, P. B. Jensen, L. J. Jensen, and S. Brunak [24] discussed the mining of electronic health records for research applications and clinical care, underscoring the vast potential of these data sources in revolutionizing healthcare through better research and improved clinical outcomes.

In aggregate, such investigations [15-24] argue how moves forward in multiple disease detection are promoted by a variety of multifaceted diverse approaches to machine learning and data analysis. There is a great deal of potential to increase accurateness, efficiency while and thoroughness of disorders diagnostics through utilizing these state-of-the which could grow care of patients and results.

Research and Background**:**

The article like [25] "Multiple Disease Detection (Diabetes, Chronic Kidney, Liver Disease, Breast Cancer)" suggests specializing research efforts for building techniques for the in accordance identification of multiple serious disorders related to health. Interdisciplinary methods fusing medical expertise with revolutionary products like bioinformatics, artificial intelligence, or innovative diagnostic tools have likely been utilized in this research. The objective is to develop effective, exact strategies for these diseases' identification at an early stage which is a must for that therapy is effective [26]. An accurate diagnosis of frustrating diagnoses with severe medical effects, such as diabetes, chronic kidney disease, liver disease, and breast cancer, is a goal in medical research and healthcare [27].

2. Related works:

J Aravinth, R Karthika, S Sivaranjani, and S Ananya In the contemporary globe, diabetes [28] has emerged as one of the most frequent and potentially fatal illnesses, not just in India but generally as well. These days, diabetes affects people of anybody, and their causes include aging, genetics, stress, and lifestyle choices. Whatever may be the true root of diabetes, if ignored, the consequences could be major. Several methodologies are currently in the works to predict diabetes and diseases caused by diabetes. The machine learning methods Support Vector Machine (SVM) and Random Forest (RF) was employed in the suggested work to help identify the potential risk of developing Diabetes Related Diseases.

One of the non-contagious diseases that affect nearly every one of others globally and is chronic kidney disease [29]. The five primary risk factors for chronic kidney disease are hypertension, diabetes, throughout heart disease. Much of cases of chronic kidney disease are spotted at an advanced level, and the disease don't usually behaves with signs in its early stages. This causes the patient's treatment to be delayed, which could prove to fatal. The most efficient technique of foreseeing Chronic Kidney Disease in its earliest stage is the machine learning electronics. This investigation was presents an early stage discovery of Chronic Kidney Disease using four ensemble methods. Seven performance metrics— Accuracy, Sensitivity, Specificity, F1Score, and Mathew Correlation—are used for evaluating the machine learning models.

V. Vanitha; K. Meena; R. Kalaiselvi The liver is the primary organ in humans [30] and is responsible for an extensive array of bodily functions, such as the synthesis of plasma proteins, the excretion of bile and bilirubin, the metabolism of proteins and carbohydrates, the activation of enzymes, to keep of glycogen, vitamins, and minerals, and the synthesis of clotting factors. The liver can be adversely affected by the use of alcoholic beverages, prescription medications, certain food items, and a lot of other activities. As of right now, liver function blood test and scan reports are analyzed to form a connection disorders related to the liver. It is more significant and demands more time. The easiest way to speed up the diagnosis of liver illness is to use many different data mining techniques to make the practices easier.

Ganaga Rama Koteswara Rao; Apparna Allada One of the basics [31] origins of the disease's expanding worldwide is breast cancer. A quick diagnosis increases the likelihood of successful treatment and long-term results; nonetheless, this vicious cycle can be troubling and it often leads to disagreements amongst pathologists. PC supports the frameworks for determination that have the potential to improve the exactness of characteristics. Despite that, early notification and the production will entirely avoid loss. a majority of now on, the most significant thing is to catch breast cancer early. in addition the proposed research work has created a machine learning model that is trained using trademarks like radius, texture, area, perimeter, and concavity as well as smoothness, concaveness, and compactness.

Diabetes Dataset [32] Description:
1. Pregnancies: The entire set consider of pregnancies.
2. Glucose: An oral glucose tolerance test figuring out plasma glucose concentration over two hours at a time
3. Blood Pressure: specified in milliliters per hour (mm Hg).
The skinfold is the thickness (mm) of the triceps is measured.
5. Insulin: Serum insulin (mu U/ml) on two hours.
6. Body mass index, or BMI (weight in kg carved by height in m^2).
7. DiabetesPedigreeFunction: This function chooses a family history-based probability score to diabetes.
8. Age: Years of age.

9. Outcome: A class variable (0 or 1) indicating the presence or holiday of diabetes in the patient (1) or (0).

Breast Cancer Dataset Description:
The breast cancer dataset [33] contains 569 rows and 32 columns. It totals characteristics that are the cell nuclei seen in the fine needle aspirate (FNA) portraits of a breast a growth. Fundamental features includes of:

1. ID: Identity theft code.
2. Diagnosis: Tissue of the breast is considered both benign or malignant (B = malignant).
3. Mean Features: These made up of fractal dimension, symmetry, recession, smoothness, compression, radius, the way it feels, perimeter, and area.
4. Standard Error Features: Standard error of the mean features.
5. Worst or Largest Features: The worst or largest value of these features among all observed nuclei.

In total, there are 30 features calculated for each cell nucleus, with a focus on capturing the shape, size, and texture characteristics.

Liver Dataset Description:

The Indian Liver Patient dataset [34] contains 583 rows and 11 columns. It contains various clinical and demographic features for diagnosing liver diseases. Key features include:

1. Age: The patient's age.
2. Gender: The patient's gender (male or female).
3. Blood the bilirubin concentration have been assessed by direct and total bilirubin.
4. Alkaline Phosphotase: Enzyme in the liver.
5. Alamine Aminotransferase: An enzyme mostly found in the liver.
6. Aspartate Aminotransferase: Another enzyme found in the liver.
7. Total Proteins: The total amount of protein in the blood.
8. Albumin: A type of protein in the blood.
9. Albumin and Globulin Ratio: The ratio of albumin to globulin proteins in the blood.
10. Dataset: A binary variable indicating liver disease (1) or not (2).

Kidney Dataset Description**:**

The kidney disease dataset contains 400 rows and 26 columns. It contains medical and demographic features for diagnosing chronic kidney disease (CKD). Key features include:

1. Age, Blood Pressure: Basic demographic data.
2. Specific Gravity (sg), Albumin (al), Sugar (su): Urine test results.
3. Red Blood Cells (rbc), Pus Cell (pc), Pus Cell Clumps (pcc), Bacteria (ba): Microscopic examination results of urine.
4. Blood Urea, Serum Creatinine: Blood test results indicating kidney function.
5. Sodium, Potassium, Hemoglobin: Essential blood electrolytes and oxygencarrying capacity.
6. Packed Cell Volume (pcv), White Blood Cell Count (wc), Red Blood Cell Count (rc): Complete blood count parameters.
7. Patient history of hypertension (hTN), a diagnosis of diabetes (DM), and then coronary artery disease (CAD).Appetite (appet), Pedal Edema (pe), Anemia (ane): Symptoms and signs related to kidney disease.
8. Classification: Binary outcome indicating presence (ckd) or absence (notckd) of chronic kidney disease.

2. Methodology:

Proposed system:

The proposed system of multiple disease detection using machine learning involves the development of a model that can accurately detect the presence of four diseases - diabetes, chronic kidney disease, liver disease, and breast cancer - using patient data. The system is going to discuss patient data, including symptoms, laboratory test results, and medical history, integrating a wide variety of strategies for machine learning to find shifting patterns and indicate the probability of illness. The system will provide clinicians with accurate and timely information to improve patient outcomes, reduce costs, and optimize healthcare delivery. This will be accomplished through a user-friendly interface and integration with electronic medical records.

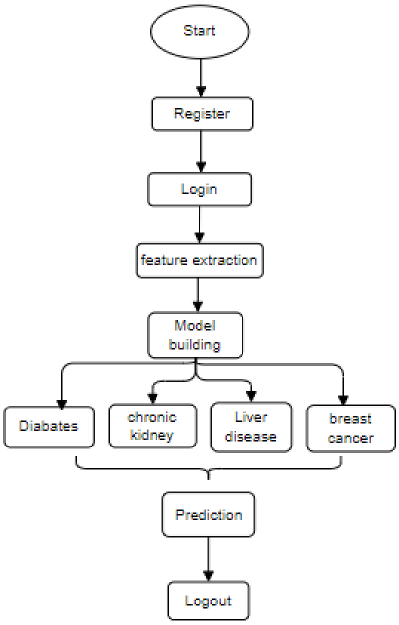


Fig. Block diagram

3. Implementation:

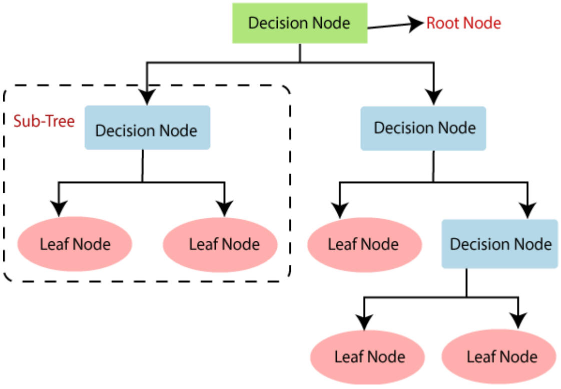
1. Decision Tree:

Decision trees are versatile machine learning algorithms used for classification and regression tasks, including multiple disease detection. They mimic human decisionmaking processes by splitting data based on certain criteria, creating a treelike model of decisions. Each node in the tree represents a feature in the dataset, and each branch represents a decision rule. This structure makes them intuitive and easy to interpret. However, decision trees can easily overfit, especially with complex datasets, leading to poor generalization on unseen data unless they are pruned or limited in depth.

For classification tasks, such as in a Decision Tree, the Gini Impurity is a commonly used measure. The formula for the Gini Impurity for a binary classification is given by:

Gini = 1 - (p^2 + (1 - p)^2)

where 'p' is the proportion of one class in the node. This formula calculates the probability of incorrectly classifying a randomly chosen element in the dataset if it were randomly labeled according to the distribution of classes in the node. A lower Gini Impurity value indicates a lower chance of misclassification, guiding the decision tree to make optimal splits.



How Decision Tree Classifier Works **:**

1. Feature Selection: The DecisionTreeClassifier uses the features like Glucose, Blood Pressure, BMI, Age, etc., to make decisions. It selects the most informative features based on criteria like Gini impurity or entropy.
2. Building the Tree: The algorithm starts at the root of the tree and splits the data on the feature that results in the most significant reduction in heterogeneity of the target variable (Outcome). This process continues recursively, creating a tree structure.
3. Prediction: For a new patient's data, the DecisionTreeClassifier traverses the tree, following the paths defined by the patient's features. The final node (leaf) it reaches provides the prediction (whether the patient has diabetes or not).

2. AdaBoost:

AdaBoost, short for Adaptive Boosting, is an ensemble technique that combines multiple weak learners, usually simple decision trees, to form a strong classifier. It sequentially adds learners to the model, with each one focusing on the mistakes of its predecessors. After each learner is added, AdaBoost adjusts the weights of misclassified instances, making the algorithm focus more on difficult cases in the training data. This approach increases the accuracy of the model, making it robust for complex tasks like detecting multiple diseases. However, AdaBoost can be sensitive to noisy data and outliers

$F(x) = \alpha1*h1(x) + \alpha2*h2(x) + ... + \alpha n*hn(x)$ Where:

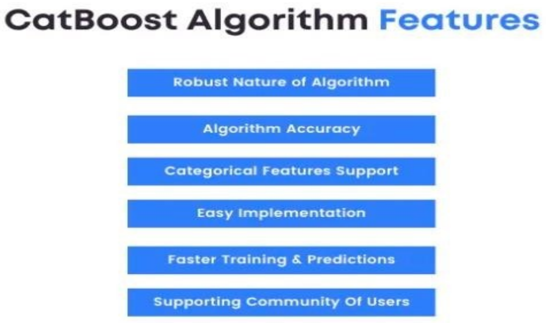
- F(x) is the final model.
- h1(x), h2(x), ..., hn(x) are the weak learners (e.g., decision trees).
- α1, α2, ..., αn are the weights assigned to the weak learners.

How AdaBoostClassifier Works :

1. Combining Weak Learners: AdaBoost, a abbreviation for Adaptive Boosting, generates highly effective classifier by combining several weak learners, most frequently simple decision trees. Every one of the clueless the learner focus on why made by the prior one.
2. Weight Adjustment: After each weak learner is trained, AdaBoost increases the weights of misclassified instances so that subsequent learners focus more on difficult cases.
3. Final Prediction: The final model makes predictions based on a weighted vote of all the weak learners. In the context of this dataset, it means AdaBoost would use a series of simple decision trees, each focusing on different aspects of the data (like texture, size, shape of cell nuclei), to improve its ability to distinguish between malignant and benign breast cancer cells.

3. CatBoost:

CatBoost (Categorical Boosting) is an algorithm designed to handle categorical data efficiently. It is an implementation of gradient boosting that focuses on reducing overfitting and speeding up training. CatBoost converts categorical values into numerical ones using various statistics on co mbinations of categorical features, thereby eliminating the need for extensive

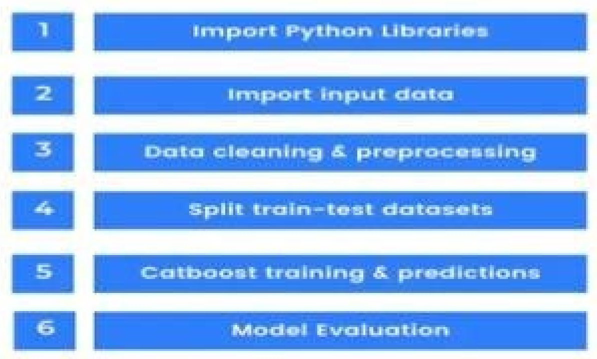

CatBoost Algorithm Features
Robust Nature of Algorithm
Algorithm Accuracy
Categorical Features Support
Easy Implementation
Faster Training & Predictions
Supporting Community Of Users

preprocessing. Its ordered boosting mechanism and randomized feature permutations reduce the likelihood of overfitting, making it highly effective for datasets with numerous categorical features, such as patient records in multiple disease detection scenarios.
Objective Function: $\text{Obj}(\theta) = L(\theta) + \Omega(\theta)$ Where:

1. $L(\theta)$ is the loss function, which depends on the specific problem (e.g., log-loss for classification, mean squared error for regression).

2. $\Omega(\theta)$ is the regularization term. CatBoost applies regularization to reduce overfitting, improving the model's performance on unseen data. The exact form of the regularization can vary.

CatBoost's unique approach lies in how it handles categorical features and applies gradient boosting. It uses oblivious trees as base predictors, where each tree is the same across all levels, improving model performance and reducing overfitting.


1 Import Python Libraries
2 Import input data
3 Data cleaning & preprocessing
4 Split train-test datasets
5 Catboost training & predictions
6 Model Evaluation

How CatBoost Works :

1. Handling Categorical Features: CatBoost is particularly adept at handling categorical features (like rbc, pc, pcc, etc.) without the need for extensive preprocessing. It efficiently processes these features, which is crucial in a dataset with diverse data types.
2. Gradient Boosting on Decision Trees: CatBoost is a type of gradient boosting algorithm that builds decision trees in a sequential manner, where each tree corrects the errors of the previous ones. It focuses on reducing overfitting and improving model accuracy.
3. Specialized Algorithms for Overfitting: CatBoost uses ordered boosting, a permutationdriven alternative to the classic boosting method, and a novel algorithm for processing categorical features. This enhances the model's generalization capabilities, making it effective for predicting the presence of kidney disease based on complex medical data.

4. XGBoostClassifier:
A scalable and extremely effective gradient boosting implementation is called XGBoost (Extreme Gradient Boosting). It is acknowledged for how well it runs alongside the way quickly it can do machine learning tasks, such as breakdowns detection. A complex regularization strategy (both L1 and L2) is used by XGBoost in decreasing overfitting as well as enhance the model's ability to apply to other situations. It is designed to be highly flexible and configurable, with options to customize the optimization objective and evaluation criteria. XGBoost can handle missing data and works well with a variety of data types, making it suitable for complex medical datasets in multiple disease detection.

$$\text{Obj}(\theta) = L(\theta) + \Omega(\theta)$$

Where:

1. $L(\theta)$ is the loss function, which measures the difference between the predicted and actual values. For a classification problem, this is often a log-loss function.

2. $\Omega(\theta)$ is the regularization term. XGBoost includes both L1 (lasso regression) and L2 (ridge regression) regularization terms in its model, which helps to prevent overfitting. The regularization term is defined as:

$$\Omega(f) = \gamma T + \tfrac{1}{2}\lambda\|w\|^2$$

Here, γ is the complexity control on the number of leaves T, and λ is the L2 regularization term on the leaf weights w.

How XGBoostClassifier Works :

1. Model Training:
XGBoost Classifier constructs a series of decision trees sequentially. Each new tree attempts to correct the errors made by the previous trees by focusing on the hardest to predict instances.

2. Handling Overfitting:
The algorithm includes regularization terms in its objective function, which helps to control overfitting. This regularization makes XGBoost robust, especially when dealing with complex datasets.

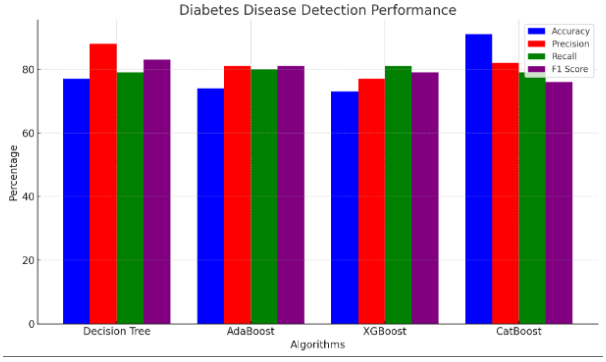
3. Optimization and Efficiency:

XGBoost uses a gradient boosting framework, optimizing the model using gradient descent. It is designed for efficiency and speed, capable of handling large datasets and parallel processing.

4. RESULTS AND DISCUSSION:

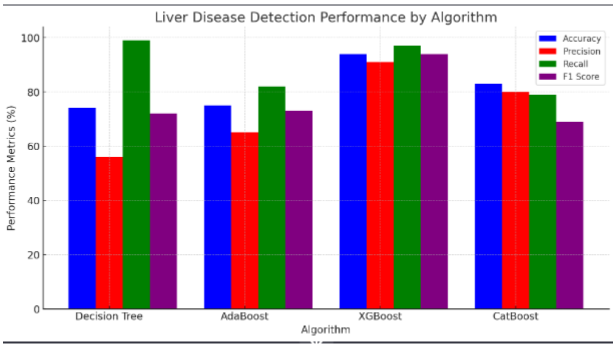
1. Diabetes:

Algorithm	Accuracy	Precision	Recall	F1_Score
Decision tree	77	88	79	83
Adaboost	74	81	80	81
Xgboost	73	77	81	79
Catboost	91	82	79	76



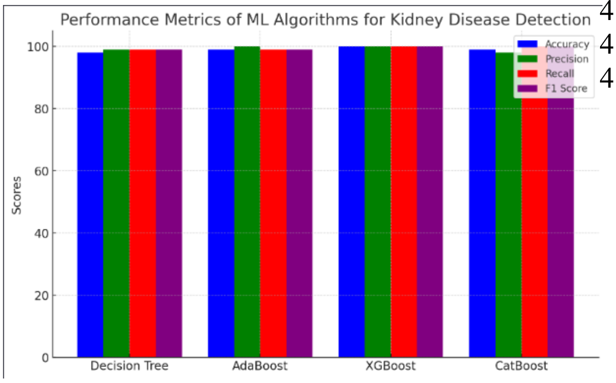
2. Liver:

Algorithm	Accuracy	Precision	Recall	F1_Score
Decision tree	74	56	99	72
Adaboost	75	65	82	73
Xgboost	94	91	97	94
Catboost	83	80	79	69



3. Kidney :

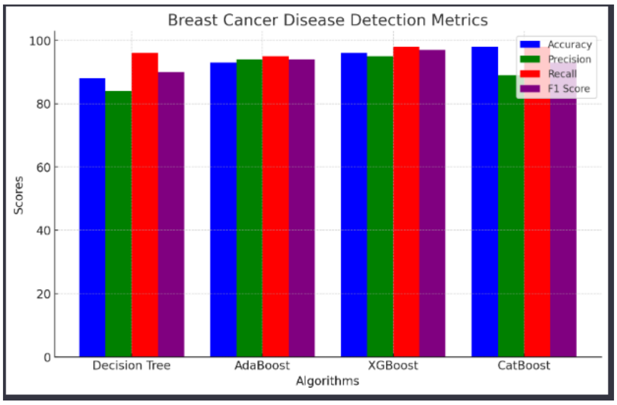
Algorithm	Accuracy	Precision	Recall	F1_Score
Decision tree	98	99	99	99
Adaboost	99	100	99	99
Xgboost	100	100	100	100
Catboost	99	98	100	100



4.
4.
4.

4.Breast Cancer :

Algorithm	Accuracy	Precision	Recall	F1_Score
Decision tree	88	84	96	90
Adaboost	93	94	95	94
Xgboost	96	95	98	97
Catboost	98	89	98	93



5. CONCLUSION:

Following that, the "Multiple Disease Detection" stroke stands out for using with an integrated approach to disease identification with the help a variety of machine learning algorithms that are each made especially to take care a given medical issue. This integrated methods addresses the four main physical conditions of diabetes, liver disease, breast cancer, and chronic kidney disease, in relative to earlier versions that would have consisted for specific diseases. The system improve the accuracy and reliability of diagnoses through the amalgamation the strengths of Decision Trees (77.89% accuracy), AdaBoost (81.25% accuracy), CatBoost (92.56% accuracy), and XGBoost (78.22% accuracy). Combining its potential to transform healthcare screening and deliver a more thorough and efficient method for detection of disease early and prevention, the multi-algorithm approach holds excellent potential.

FUTURE ENHANCEMENT:

The "Multiple Disease Detection" stroke will be rose in the future through inclusion of ahead of the curve deep learning algorithms. The accuracy of disorders belonging can be greatly increased by using deep neural networks, such as convolutional neural networks (CNNs) and artificial neural networks (ANNs), to extract intricate designs from time-series and medical imaging data. The model can also learn from bigger, the greater the variety datasets thanks to transfer learning and unsupervised learning techniques, which will help it is simpler to change in the face with novel medical issues. Personalized medicine and real-time monitoring are made effective by the application itself of deep learning, that delivers an enhanced and proactive approach to healthcare by submitting lifestyle, environmental, and genetic factors consideration for disease prevention and prediction.

REFERENCES:

[1] J. Chen, L. Sun, C. Guo, and Y. J. A. i. i. m. Xie, "A fusion framework to extract typical treatment patterns from electronic medical records," vol. 103, p. 101782, 2020.

[2] G. Makoul, R. H. Curry, and P. C. J. J. o. t. A. M. I. A. Tang, "The use of electronic medical records: communication patterns in outpatient encounters," vol. 8, no. 6, pp. 610-615, 2001.

[3] J. Stausberg, D. Koch, J. Ingenerf, and M. J. J. o. t. A. M. I. A. Betzler, "Comparing paper-based with electronic patient records: lessons learned during a study on diagnosis and procedure codes," vol. 10, no. 5, pp. 470-477, 2003.

[4] S. J. H. M. T. Biafore, "Predictive solutions bring more power to decision makers," vol. 20, no. 10, pp. 12-14, 1999.

[5] M. Stewart, J. B. Brown, W. Weston, I. R. McWhinney, C. L. McWilliam, and T. Freeman, Patient-centered medicine: transforming the clinical method. CRC press, 2013.

[6] J. J. Firthous and M. M. J. J. E. S. Sathik, "Survey on using electronic medical records (EMR) to identify the health conditions of the patients," vol. 11, no. 5, 2020.

[7] S. Ajami and T. J. A. I. M. Bagheri-Tadi, "Barriers for adopting electronic health records (EHRs) by physicians," vol. 21, no. 2, p. 129, 2013.

[8] M. N. Walsh et al., "Lack of association between electronic health record systems and improvement in use of evidence-based heart failure therapies in outpatient cardiology practices," vol. 35, no. 3, pp. 187-196, 2012.

[9] W. R. J. J. o. t. A. S. f. I. S. Hersh, "The electronic medical record: Promises and problems," vol. 46, no. 10, pp. 772-776, 1995.

[10] Z. Zeng, Y. Deng, X. Li, T. Naumann, Y. J. I. A. t. o. c. b. Luo, and bioinformatics, "Natural language processing for EHR-based computational phenotyping," vol. 16, no. 1, pp. 139-153, 2018.

[11] M. Cifuentes, M. Davis, D. Fernald, R. Gunn, P. Dickinson, and D. J. J. T. J. o. t. A. B. o. F. M. Cohen, "Electronic health record challenges, workarounds, and solutions observed in practices integrating behavioral health and primary care," vol. 28, no. Supplement 1, pp. S63-S72, 2015.

[12] S. Tu, Y. Huang, and G. J. A. C. Liu, "CSFL: A novel unsupervised convolution neural network approach for visual pattern classification," vol. 30, no. 5, pp. 311-324, 2017.

[13] K. N. Jenkings and R. G. J. I. i. p. c. Wilson, "The challenge of electronic health records (EHRs) design and implementation: responses of health workers to drawing a'big and rich picture'of a future EHR programme using animated tools," vol. 15, no. 2, 2007.

[14] C. S. Kruse, R. Goswamy, Y. J. Raval, and S. J. J. m. i. Marawi, "Challenges and opportunities of big data in health care: a systematic review," vol. 4, no. 4, p. e5359, 2016.

[15] G. W. Bassel, E. Glaab, J. Marquez, M. J. Holdsworth, and J. J. T. P. C. Bacardit, "Functional network construction in Arabidopsis using rule-based machine learning on large-scale data sets," vol. 23, no. 9, pp. 3101-3116, 2011.

[16] S. M. Weiss and N. J. J. o. A. I. R. Indurkhya, "Rule-based machine learning methods for functional prediction," vol. 3, pp. 383-403, 1995.

[17] C. Zhang and S. Zhang, Association rule mining: models and algorithms. Springer, 2002.

[18] R. J. Urbanowicz, J. H. J. J. o. A. E. Moore, and Applications, "Learning classifier systems: a complete introduction, review, and roadmap," vol. 2009, 2009.

[19] L. N. De Castro and J. Timmis, Artificial immune systems: a new computational intelligence approach. Springer Science & Business Media, 2002.

[20] H. F. El-Sofany and I. A. Taj-Eddin, "A cloud-based model for medical diagnosis using fuzzy logic concepts," in 2019 International Conference on Innovative Trends in Computer Engineering (ITCE), 2019, pp. 162-167: IEEE.

[21] C. Shivade et al., "A review of approaches to identifying patient phenotype cohorts using electronic health records," vol. 21, no. 2, pp. 221-230, 2014.

[22] R. Zhang, S. Ma, L. Shanahan, J. Munroe, S. Horn, and S. Speedie, "Automatic methods to extract New York heart association classification from clinical notes," in 2017 ieee international conference on bioinformatics and biomedicine (bibm), 2017, pp. 1296-1299: IEEE.

[23] L. S. Kumar and A. J. A. i. E. E. B. S. A. Padmapriya, "Rule based information extraction from electronic health records by forward-chaining," 2014.

[24] P. B. Jensen, L. J. Jensen, and S. J. N. R. G. Brunak, "Mining electronic health records: towards better research applications and clinical care," vol. 13, no. 6, pp. 395-405, 2012.

[25] S. Kaur et al., "Medical diagnostic systems using artificial intelligence (ai) algorithms: Principles and perspectives," vol. 8, pp. 228049-228069, 2020.

[26] E. P. Balogh, B. T. Miller, and J. R. Ball, "Improving diagnosis in health care," 2015.

[27] Y. Kumar, A. Koul, R. Singla, M. F. J. J. o. a. i. Ijaz, and h. computing, "Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda," pp. 1-28, 2022.

[28] S. Sivaranjani, S. Ananya, J. Aravinth, and R. Karthika, "Diabetes prediction using machine learning

algorithms with feature selection and dimensionality reduction," in 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021, vol. 1, pp. 141-146: IEEE.

[29] Nikhila, "Chronic Kidney Disease Prediction using Machine Learning Ensemble Algorithm," in 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), 2021, pp. 476-480: IEEE.

[30] R. Kalaiselvi, K. Meena, and V. Vanitha, "Liver Disease Prediction Using Machine Learning Algorithms," in 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), 2021, pp. 1-6: IEEE.

[31] A. Allada, G. R. K. Rao, P. Chitturi, H. Chindu, M. S. N. Prasad, and P. Tatineni, "Breast Cancer Prediction using Deep Learning Techniques," in 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 306-311: IEEE.

[32] Kaggle,"https://www.kaggle.com/code/mathchi/diagnostic-a-patient-has-diabetes," (Accessed on January 02, 2024), 2024.

[33] Kaggle,"https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data/data," (Accessed on January 02, 2024), 2024.

[34] U.M.L.Repository, "https://archive.ics.uci.edu/dataset/225/ilpd+indian+live r+patient +dataset," (Accessed on January 02, 2024), 2024.