Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 5s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Data Mining-Driven Multi-Feature Selection for Chronic Disease Forecasting

Dr.B Rama Ganesh[1], Praveen B M[2], Krishna Prasad K[3], G.Swapna[4], Viswanath G[5]

[1]Post-Doctoral Fellow, Srinivas University, Mangalore, Karnataka, India.
Email ID: ramaganesh34@gmail.com
[2]Professor, Institute of Engineering and Technology, Srinivas University, Mukka-574146, Karnataka, India.
ORCID-ID: 0000-0003-2895-5952
Email ID: bm.praveen@yahoo.co.in
[3]Professors, Institute of Engineering and Technology, Srinivas University, Mukka-574146, Karnataka, India.
ORCID-ID: 0000-0001-5282-9038
Email ID: krishnaprasadkcci@srinivasuniversity.edu.in
[4]Assistant Professor, Apollo institute of pharmaceutical sciences, The Apollo University, Chittoor, India.
ORCID-ID: 0000-0002-9340-4148
Email ID: swapnagv111@gmail.com
[5]Associate Professor, Department of CSE-AIML, Sri Venkatesa Perumal College of Engineering and Technology, Puttur, India.
ORCID-ID: 0009-0001-7822-4739
Email ID: viswag111@gmail.com

Cite this paper as: Dr.B Rama Ganesh, Praveen B M, Krishna Prasad K, G.Swapna, Viswanath G, (2025) Data Mining-Driven Multi-Feature Selection for Chronic Disease Forecasting. *Journal of Neonatal Surgery*, 14 (5s), 108-124.

ABSTRACT
a major worldwide health difficulty, continual sicknesses call for higher predictive fashions for early diagnosis and individualized treatment. the usage of a synergistic blend of strategies, this technique combines Recursive characteristic removal with pass-Validation (RFECV) and support Vector machine (SVM) for best characteristic selection throughout 8 wonderful datasets: Breast cancer, chronic Kidney, Diabetes danger, Erbil heart sickness, coronary heart disorder, Kidney disease, Pima Indians, and Wisconsin Breast. The technique stresses efficient dimensionality reduction so that the most pertinent facts is applied to improve model overall performance. With a voting Classifier combining AdaBoost decision Tree and ExtraTree obtaining outstanding performance across all datasets, ensemble learning is absolutely important. High-performance results from this era show its dependability and relevance for early continual disorder prediction. The method provides brilliant possibility for enhancing diagnostic accuracy and allowing spark off remedies by tackling feature selection issues and imposing ensemble learning, thereby enhancing healthcare management and results.

Keywords: *Synergistic Feature Engineering, chronic disease, early prediction, machine learning, Voting Classifier, Ensemble learning.*

1. INTRODUCTION

With ailments including cardiovascular illnesses, cancer, kidney sickness, and diabetes many of the maximum not unusual and sizeable worldwide incidence of continual sicknesses, the healthcare field is severely concerned about them. together with causing high-quality morbidity and demise, those illnesses severely tax healthcare systems and society at large financially [1]. The load of continual illnesses is anticipated to growth as the arena populace a long time and way of life-associated danger elements grow to be extra not unusual; so, innovative ideas in healthcare are alternatively vital [2]. Especially in their early degrees, the continual character of many diseases method they often development silently; issues encompass coronary heart ailment, diabetes, and several malignancies live asymptomatic till they have got advanced stages. [28] Early detection is particularly crucial as this not on time identification often causes irreparable consequences [3].

Preventing the spread of persistent sicknesses and reducing associated fitness dangers rely on early degree identification and intervention capabilities for them. Early identification not handiest will increase analysis but also makes it feasible to apply

customized treatments intended for precise risk profiles. This early cognizance can help to significantly decrease medical resource load and healthcare charges. The early analysis of persistent illnesses has advanced substantially with the appearance of artificial intelligence (AI), for this reason enhancing the ability of medical practitioners to identify excessive-chance individuals at a younger stage[27]. mainly machine learning (ML), artificial intelligence (AI) technologies have shown extraordinary promise in comparing complicated medical data, recognizing tendencies, and producing predictions impossible for human experts to look [4]. AI models can supply tailored forecasts with the aid of regular studying and data model, consequently assisting healthcare clinicians in growing extra precise remedy regimens relying on man or woman hazard variables [8].

By means of synthetic intelligence integration into healthcare systems, chronic disorder control provides a remodeling answer that enhances prognosis accuracy, treatment strategy optimization, and ultimately patient outcomes [5]. By way of their capacity to method giant quantities of clinical data, machine learning algorithms were efficaciously applied in numerous domains together with early disorder identification, enhancement of diagnostic accuracy, and disorder development prediction [9]. the search of machine mastering methods is steadily helping to struggle continual sicknesses with the exponential [29] increase in clinical records availability since it provides more regular and efficient gadgets for healthcare practitioners to paintings upon. This invention is commencing the path for a time whilst artificial intelligence now not handiest supports prognosis but also customised healthcare interventions, so helping to lower the worldwide impact of continual sicknesses [10], [11].

2. RELATED WORK

Studies on early identification and treatment of persistent illnesses such as heart ailment, diabetes, most cancers, and renal sickness has been prompted through their mounting load on international healthcare systems. specializing in growing the accuracy of forecasts and making an allowance for early responses, numerous studies have addressed those difficulties utilizing machine learning approaches. Deep learning, ensemble learning, and feature selection among different tactics have been investigated to enhance the prediction version performance in analysis of persistent diseases.

Combining XGBoost with a deep neural network (DNN), Maleki et al. [12] suggested to diagnose breast most cancers from histomorphology images. Promising results within the identification of malignant tissues were received by this hybrid strategy the usage of XGBoost for category and deep learning for feature extraction: In identical vein, Raihan et al. [13] identified chronic kidney disease (CKD) using the XGBoost classifier. To similarly openness and understanding of the selection-making process, in addition they used SHAP (Shapley Additive motives) to investigate how extraordinary aspects affect the version. This method underlined the want of particular factors, including creatinine and urea ranges, in CKD analysis, so strengthening version reliability.

Hegde and Mundada [14] supplied an adaptive probabilistic divergence-based feature choice method inside the field of chronic sickness prediction to elevate the machine learning algorithm accuracy. This method sought to reduce dimensionality at the same time as maintaining pertinent information through choosing the maximum informative functions from healthcare data. Their examine showed how well probabilistic divergence might be used to maximize characteristic selection and thereby machine learning version performance. Maini et al. [15] investigated stacking generalization in addition to be able to improve persistent ailment prediction system performance. by using the use of their combined strengths, they shaped an ensemble from several machine learning models, therefore mitigating the restrictions of individual algorithms. relatively to single-model techniques, the stacked generalization approach proved to be rather successful and accelerated prediction accuracy.

Additional studies have looked at how tender clustering might help to diagnose chronic diseases more correctly. Aldhyani et al. [16] provided a smooth clustering technique and combined it with machine learning techniques to pick out chronic sicknesses. greater flexible data classification made possible by using soft clustering helped the system to manage not unusual in scientific analysis difficult conditions. particularly inside the context of complicated and noisy healthcare data, this has proven potential in elevating diagnostic accuracy.

Through an all-encompassing approach, Jongbo et al. [17] also helped to diagnose continual renal ailment. Emphasizing the need of ensemble methods in healthcare prediction sports, their version included several classifiers to beautify diagnostic performance. Their method attained higher accuracy, robustness, and dependability by using inclusive of more fashions than by the use of unmarried classifiers. Likewise, Listiana et al. [18] used information gain and AdaBoost to enhance the accuracy of an optimal “support vector machine (SVM)” for the detection of chronic kidney disease. Their technique confirmed the fee of combining several optimization techniques in the prediction of persistent diseases by the use of SVM's power in managing excessive-dimensional data to enhance the performance of the classifier using AdaBoost.

other studies have targeting enhancing prediction model explainability and interpretability. for instance, the use of SHAP values in Raihan et al. [13] study permit doctors to grasp the particular contributions of each feature within the prediction technique, so improving self belief in the results of the version. In healthcare programs, in which practitioners' adoption of models may be a whole lot motivated by their arrival at their conclusions, this awareness on explainability is simply

important.

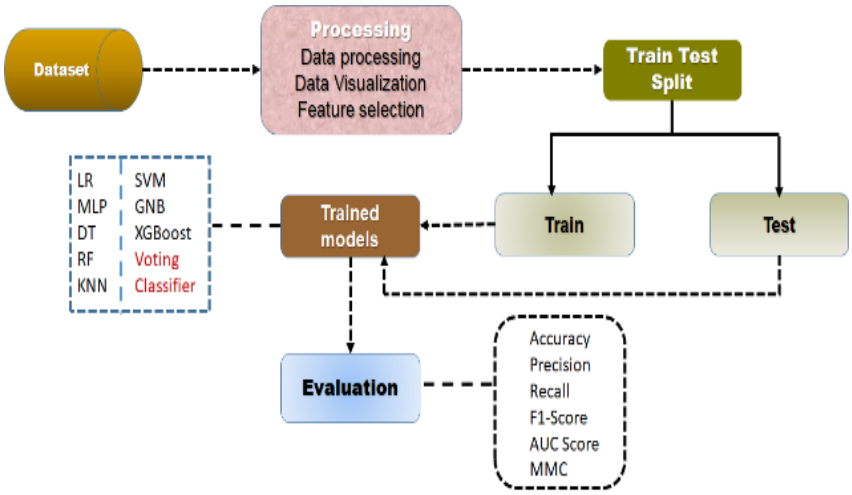
The literature frequently unearths a ordinary motif in the integration of ensemble strategies in continual contamination prediction: Predictive overall performance has been confirmed to be advanced by stacked generalization, smooth grouping, and the mix of several classifiers. Combining the strengths of numerous algorithms facilitates these methods resolve model overfitting, decorate generalization, and raise robustness. In healthcare, in which the complexity of medical data needs models that can control several data types and offer correct effects throughout a huge variety of affected person demographics and conditions, ensemble learning models—like those proposed through Maini et al. [15] and Jongbo et al. [17] are specifically treasured.

Regarding characteristic selection, methods consisting of information benefit [18] and probabilistic divergence [14] have proven success in pointing up the maximum pertinent elements from large datasets. those processes no longer simplest enhance version overall performance but also guarantee that the produced fashions are more interpretable and computationally low-priced by reducing the dimensionality of the data while maintaining vital information.

furthermore, traits in deep learning have greatly improved illness identification from intricate data assets like scientific imaging. Maleki et al. [12] confirmed how well deep learning mixed with traditional machine learning methods—which includes XGBoost—may growth diagnostic accuracy. Deep learning shines in characteristic extraction, for this reason these hybrid models are especially useful for handling high-dimensional and unstructured data such as histopathology images.

3. MATERIALS AND METHODS

using advanced machine learning algorithms and characteristic engineering techniques throughout eight datasets—Breast cancer [19], chronic Kidney [20], Diabetes risk [21], Erbil heart disease [22], coronary heart disorder [23], Kidney disorder [24], Pima Indians [25], and Wisconsin Breast [26] the proposed device makes a speciality of early prediction of persistent diseases. Recursive characteristic elimination [7] with cross-Validation [RFECV] and "support Vector machine (SVM)" is applied characteristic selection to locate the most applicable predictors. Logistic Regression, Multilayer Perceptron (MLP), "decision Tree, Random forest, k-Nearest neighbors (KNN), support Vector machine (SVM), Gaussian Naive Bayes" (GaussianNB), and XGBoost [13] improved through Bayesian Optimization is among the thorough array of strategies the system combines. to improve predictive overall performance, a sturdy ensemble technique is used with a voting Classifier combining AdaBoost decision Tree with ExtraTree. The system seeks to provide a scalable and efficient solution for early continual sickness identification by combining feature selection with other algorithms, therefore permitting timely treatments and better healthcare effects.



"Fig.1 Proposed Architecture"

The system architecture (fig. 1) consists in a series of movements to evaluate performance, educate models, and deal with data. records processing and visualization start it; [30] function selection comes subsequent. training and test sets then separate the dataset. training the use of the schooling records a couple of fashions (LR, SVM [18], MLP, GNB, DT, XGBoost [12], RF, KNN, voting Classifier) are finished. Following criteria which includes "accuracy, precision, recall, F1-score, AUC score, and MMC," the skilled models are then examined on the testing data.

i) Dataset Collection:

Comprising 569 entries and 14 features—consisting of both feature columns and a goal variable—the Breast cancer dataset [19] Represented as an integer (0 for benign, 1 for malignant), the functions are "radius_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, symmetry_mean, texture_se, radius_worst, smoothness_worst,

compactness_worst, concavity_worst, concave points_worst, symmetry_worst," target variable diagnosis. each first-class are numerical; there are no lacking values.

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness
0	842302	M	17.99	10.38	122.80	1001.0	(
1	842517	M	20.57	17.77	132.90	1326.0	(
2	84300903	M	19.69	21.25	130.00	1203.0	(
3	84348301	M	11.42	20.38	77.58	386.1	(
4	84358402	M	20.29	14.34	135.10	1297.0	(

"Fig.2 Dataset Collection for Breast Cancer"

Comprising 491 entries and 12 attributes, the chronic Kidney dataset [20] those traits incorporate: sex, ACEIARB (Angiotensin changing Enzyme Inhibitors or Angiotensin Receptor Blockers), TIME_YEAR, and the goal variable EventCKD35, which shows the occurrence of chronic Kidney disease (CKD), HistoryDiabetes, HistoryCHD, HistoryVascular, HistorySmoking, HistoryHTN, HistoryDLD (Dyslipidemia), DMmeds, HTNmeds each quality suit either binary or specific patterns.

	Sex	AgeBaseline	HistoryDiabetes	HistoryCHD	HistoryVascular	HistorySmoking	HistoryH
0	0	64	0	0	0	0	
1	0	52	0	0	0	0	
2	0	56	0	0	0	0	
3	0	58	0	0	0	0	
4	0	63	1	0	0	0	

"Fig.3 Dataset Collection for Chronic kidney"

Comprising 144 entries and 7 attributes, the Diabetes risk dataset [21] these characteristics comprise gender, "polyuria, polydipsia, sudden weight loss, genital thrush, irritability," and the goal variable class, consequently indicating the presence of diabetes danger. every nice are binary or categorical; values are expressed as numbers. Having no missing values, the dataset gives insights on factors inflicting diabetes danger.

	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itchir
0	40	Male	No	Yes	No	Yes	No	No	No	Y
1	58	Male	No	No	No	Yes	No	No	Yes	N
2	41	Male	Yes	No	No	Yes	Yes	No	No	Y
3	45	Male	No	No	Yes	Yes	Yes	Yes	No	Y
4	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Y

"Fig.4 Dataset Collection for Diabetes Risk"

With 333 items and 3 attributes—ecgpatt, which represents ECG pattern data, qwave, indicating the life of a Q-wave in the ECG, and target, the binary target variable indicating either presence or absence of heart sickness—the Erbil heart sickness dataset [22] every fine are numerical without any missing values.

	age	sex	smoke	years	ldl	chp	height	weight	fh	active	...	ihd	hr	dm	bpsys
0	65	0	0	0	69.0	4	168	111.0	1	0	...	1	98	1	120
1	54	1	0	0	117.0	2	145	81.0	0	0	...	0	85	0	130
2	61	0	1	45	86.2	2	160	72.0	0	0	...	0	63	1	150
3	57	0	0	0	76.0	2	176	78.0	1	0	...	1	74	1	120
4	62	1	0	0	160.0	3	154	61.0	0	0	...	0	89	1	110

"Fig.5 Dataset Collection for Erbil Heart Disease"

With 1025 entries and 12 attributes—sex, cp "(type of chest pain), trestbps (resting blood pressure), fbs (fasting blood sugar), restecg (resting electrocardiogram), thalach (maximum coronary heart fee executed), exang (workout induced angina), oldpeak (depression induced via exercise), slope (slope of the peak exercise ST segment), ca (number of main vessels coloured via fluoroscopy), target (present or absent heart infection) [23] and thal (thalassemia)."

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	targe
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	

"Fig.6 Dataset Collection for Heart Disease"

There are 158 objects in the dataset with 3 attributes: al (albumin levels), sc (serum creatinine levels), and class—whether the patient has renal disorder, expressed as a binary cost. based on those essential medical factors, renal infection is predicted from the dataset [24].

	id	age	bp	sg	al	su	rbc	pc	pcc	ba	...	pcv	wc	rc
0	0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	...	44	7800	5.2
1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	...	38	6000	NaN
2	2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	...	31	7500	NaN
3	3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	...	32	6700	3.9
4	4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	...	35	7300	4.6

"Fig.7 Dataset Collection for kidney Disease"

393 entries totaling 8 attributes—Pregnancies, Glucose, "BloodPressure, Insulin, BMI, DiabetesPedigreeFunction," Age, and outcome—make up the Pima Indians Diabetes dataset [25]. based on medical traits like glucose levels, insulin, BMI, and other health signs, this dataset is used to forecast the diabetes chance in Pima Indian women.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFuncti
0	6	148	72	35	0	33.6	0.6
1	1	85	66	29	0	26.6	0.3
2	8	183	64	0	0	23.3	0.6
3	1	89	66	23	94	28.1	0.1
4	0	137	40	35	168	43.1	2.2

"Fig.8 Dataset Collection for Pima Indians"

Six thousand items with seven attributes make up the Wisconsin Breast cancer dataset [31]: clump_thickness, shape_uniformity, marginal_adhesion, bare_nucleoli, bland_chromatin, mitoses, and class. these properties outline certain aspects of the mobile nuclei seen in breast cancer biopsies; the dataset is used to categorize tumors as benign or malignant depending on those criteria.

	id	clump_thickness	size_uniformity	shape_uniformity	marginal_adhesion	epithelial_
0	1000025	5	1	1	1	
1	1002945	5	4	4	5	
2	1015425	3	1	1	1	
3	1016277	6	8	8	1	
4	1017023	4	1	1	3	

"Fig.9 Dataset Collection for Wisconsin Breast"

ii) Pre-Processing:

Our pre-processing concentration is on getting the data ready for modelling. To assure first-class of input for the prediction model, this covers data processing, showing vital correlations, and feature selection.

a) Data Processing: data processing includes handling missing values, duplication removal, and encoding of categorical variables so cleaning the dataset. [26] Standardizing numerical values by feature scaling ensures models run as anticipated. schooling and trying out units separate the facts for version evaluation. these preprocessing moves improve the excellent of the data thereby facilitating powerful model education and lowering of prediction bias.

b) Data Visualization*:* data visualization aids in the discovery of dataset correlations and tendencies. Multi-collinearity is determined via visualizing relationships among numerical features using a correlation matrix. [32] furthermore, a pattern outcome visualization indicates the goal variable's distribution, so assisting to grasp magnificence imbalance or distribution in class problems.

c) Feature Selection: using cross-Validation (RFECV), recursive characteristic removal [7] is carried out for characteristic selection, thereby iteratively eliminating less essential ones to discover the maximum vital ones. [33] using a support Vector machine (SVM) model, RFECV unearths the satisfactory subset of capabilities relying on performance, therefore lowering overfitting and raising version accuracy.

iii) Training & Testing:

usually utilizing an 80-20 or 70-30 ratio, the dataset is cut up into education and trying out sets, consequently making sure that the model is trained on maximum of the data and retains a few for evaluation. [34] The machine learning version is advanced using the schooling set, consequently maximizing the parameters of the version. The generalizing potential of the version is evaluated on the trying out set following training. version performance is expressed using metrics like accuracy, precision, recall, and F1 score.

iv) Algorithms:

LR: Binary classification problems—especially when the outcome is a binary result—use logistic regression. it is ideal for

estimating the chance of sophistication club in distinct clinical databases since it uses a logistic function to simulate the link between the dependent and independent variables.

MLP: difficult, non-linear interactions between features are captured using multilayer perceptron. Its several layers of neurons allow the model to learn complex patterns within the facts and perform well in type problems with significant characteristic areas.

DT: data is classified depending on feature thresholds using a decision tree. Specifically for express outputs, it reduces impurity by recursively partitioning the dataset depending on feature values, therefore imparting interpretable decision-making principles and notable performance on based datasets.

RF: by averaging predictions from numerous decision trees, a Random forest ensemble method is used to elevate class accuracy. [35] It is resistive to noise, works nicely on many datasets, and manages overfitting better than individual decision trees, so offering strong predictions for clinical prognosis applications.

KNN: finding the most often taking place elegance among the closest acquaintances of a data item helps k-Nearest pals to classify facts. For smaller datasets, this non-parametric technique is easy but green as, in classification troubles, similarity among data factors offers clear decision-making.

SVM: assurance finding the best hyperplane separating numerous classes with the biggest margin allows to categorise data the usage of vector machines. Particularly useful in excessive-dimensional environments, it [18] plays properly for packages desiring unique class separation, such threat prediction or disorder type.

GNB: category problems involving assumed Gaussian distribution of features use Gaussian Naive Bayes. when the dataset consists of non-stop variables and provides probabilistic forecasts for the likelihood of class club, it's miles in particular beneficial as easy, powerful device.

XGB: by iteratively improving hyperparameters, XGBoost in concert with Bayesian optimization improves predictive performance. For big datasets, it [6] is quite a hit since it generates sturdy, correct models less liable to overfitting that is essential for hard classification tasks.

Voting Classifier: Combining the predictions from several classifiers—including AdaBoost [18] and ExtraTree—is executed through a balloting classifier. This mixed technique uses the talents of several fashions to elevate standard accuracy, therefore producing a more dependable and robust category result.

4. RESULTS & DISCUSSION

Accuracy: The capacity of a test to accurately distinguish between patients and healthy individuals defines its accuracy. The calculation of the ratio of true positives and true negatives among all studied cases will enhance the accuracy of the assessment. It is articulated mathematically:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} (1)$$

Precision: Precise measurements of positivity categorize positivity along with components or portions of samples. The formula for calculating accuracy is as follows:

$$Precision = \frac{\text{True Positive}}{\text{True Positive } + \text{False Positive}} (2)$$

Recall: In machine learning, recall is a statistic gauging a version's capacity to find all pertinent times of a given magnificence. It gives information at the completeness of a version in phrases of accurately predicted positive observations to the general real positives.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} (3)$$

F1-Score: In machine learning, The F1 point model is a metric of purity. It integrates a model's memory with precise ranking. Throughout the dataset, the accuracy degree counts the variety of instances a model produced a proper prediction.

$$F1\ Score = 2 * \frac{Recall\ \mathrm{X}\ Precision}{Recall + Precision} * 100 (1)$$

AUC-ROC Curve: A performance evaluation for class problems at numerous threshold values is the AUC-ROC Curve. Plotting the authentic advantageous fee towards the false effective rate, ROC in which a better AUC denotes higher model performance, AUC measures the overall capacity of the version to distinguish between lessons.

$$AUC = \sum_{i=1}^{n-1} (FPR_{i+1} - FPR_i) \cdot \frac{TPR_{i+1} + TPR_i}{2} (5)$$

MCC: In machine learning, the Matthews coefficient—also called the "Matthews correlation coefficient (MCC—performance measure for binary classifiers)". with regard to all four components of a confusion matrix, it gauges the relationship between the predicted and real binary effects.

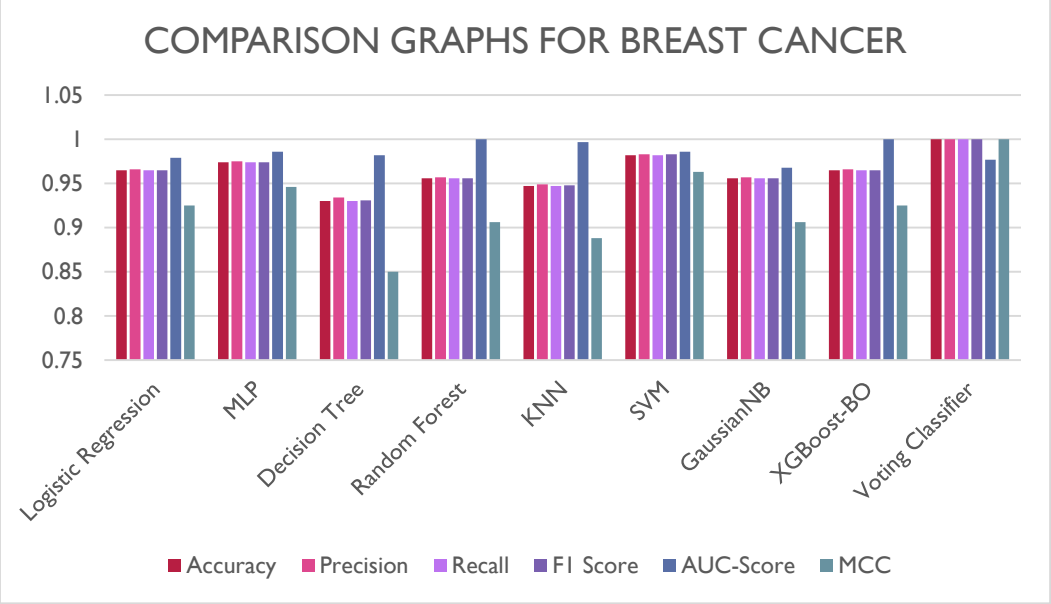
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN}} \quad (6)$$

Tables (1 to 8) analyze for each algorithm the performance measures: accuracy, precision, recall, F1-score, AUC-score, and MCC. The voting Classifier routinely beats all other methods over all datasets. Furthermore providing a comparative study of the metrics for the various algorithms are the tables.

"Table.1 Performance Evaluation Metrics for Breast Cancer"

ML Model	Accuracy	Precision	Recall	F1 Score	AUC-Score	MCC
Logistic Regression	0.965	0.966	0.965	0.965	0.979	0.925
MLP	0.974	0.975	0.974	0.974	0.986	0.946
Decision Tree	0.930	0.934	0.930	0.931	0.982	0.850
Random Forest	0.956	0.957	0.956	0.956	1.000	0.906
KNN	0.947	0.949	0.947	0.948	0.997	0.888
SVM	0.982	0.983	0.982	0.983	0.986	0.963
GaussianNB	0.956	0.957	0.956	0.956	0.968	0.906
XGBoost-BO	0.965	0.966	0.965	0.965	1.000	0.925
Voting Classifier	1.000	1.000	1.000	1.000	0.977	1.000

"Graph.1 Comparison Graphs for Breast Cancer"

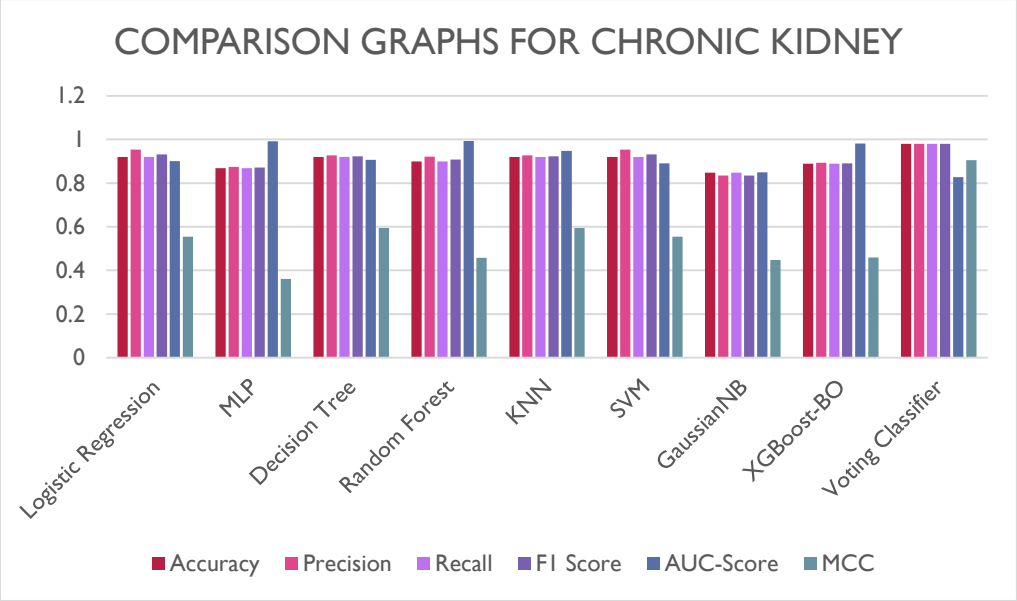


"Table.2 Performance Evaluation Metrics for Chronic Kidney"

ML Model	Accuracy	Precision	Recall	F1 Score	AUC-Score	MCC

Logistic Regression	0.919	0.954	0.919	0.931	0.901	0.554
MLP	0.869	0.874	0.869	0.871	0.991	0.361
Decision Tree	0.919	0.927	0.919	0.922	0.907	0.594
Random Forest	0.899	0.921	0.899	0.908	0.993	0.458
KNN	0.919	0.927	0.919	0.922	0.947	0.594
SVM	0.919	0.954	0.919	0.931	0.890	0.554
GaussianNB	0.848	0.834	0.848	0.835	0.849	0.448
XGBoost-BO	0.889	0.893	0.889	0.891	0.981	0.460
Voting Classifier	0.980	0.980	0.980	0.980	0.828	0.905

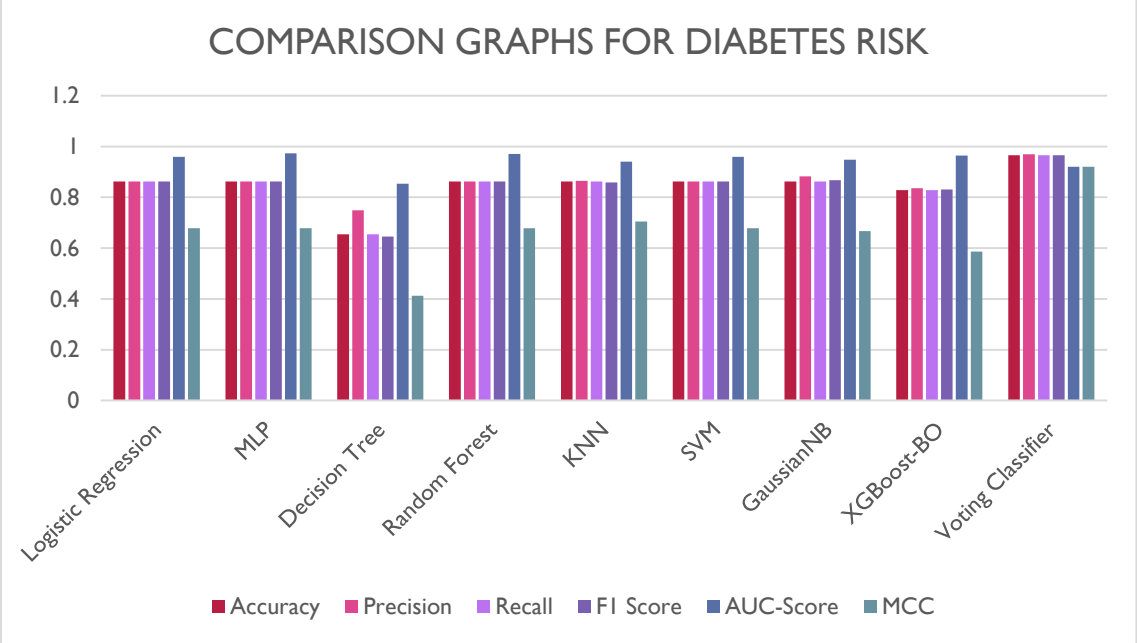
"Graph.2 Comparison Graphs for Chronic Kidney"



"Table.3 Performance Evaluation Metrics for Diabetes Risk"

ML Model	Accuracy	Precision	Recall	F1 Score	AUC-Score	MCC
Logistic Regression	0.862	0.862	0.862	0.862	0.959	0.678
MLP	0.862	0.862	0.862	0.862	0.973	0.678
Decision Tree	0.655	0.749	0.655	0.645	0.854	0.412
Random Forest	0.862	0.862	0.862	0.862	0.971	0.678
KNN	0.862	0.865	0.862	0.859	0.940	0.705
SVM	0.862	0.862	0.862	0.862	0.959	0.678
GaussianNB	0.862	0.882	0.862	0.867	0.948	0.667
XGBoost-BO	0.828	0.836	0.828	0.831	0.964	0.587
Voting Classifier	0.966	0.969	0.966	0.966	0.920	0.920

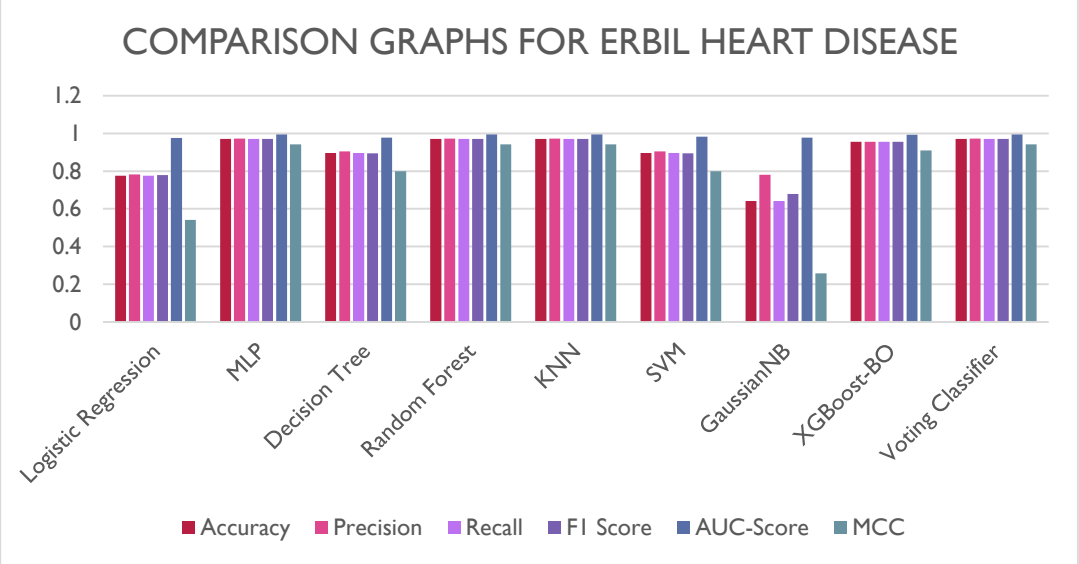
"Graph.3 Comparison Graphs for Diabetes Risk"



"Table.4 Performance Evaluation Metrics for Erbil Heart Disease"

ML Model	Accuracy	Precision	Recall	F1 Score	AUC-Score	MCC
Logistic Regression	0.776	0.783	0.776	0.778	0.975	0.541
MLP	0.970	0.972	0.970	0.970	0.995	0.941
Decision Tree	0.896	0.905	0.896	0.895	0.977	0.800
Random Forest	0.970	0.972	0.970	0.970	0.995	0.941
KNN	0.970	0.972	0.970	0.970	0.994	0.941
SVM	0.896	0.905	0.896	0.895	0.983	0.800
GaussianNB	0.642	0.781	0.642	0.679	0.977	0.257
XGBoost-BO	0.955	0.955	0.955	0.955	0.993	0.910
Voting Classifier	0.970	0.972	0.970	0.970	0.995	0.941

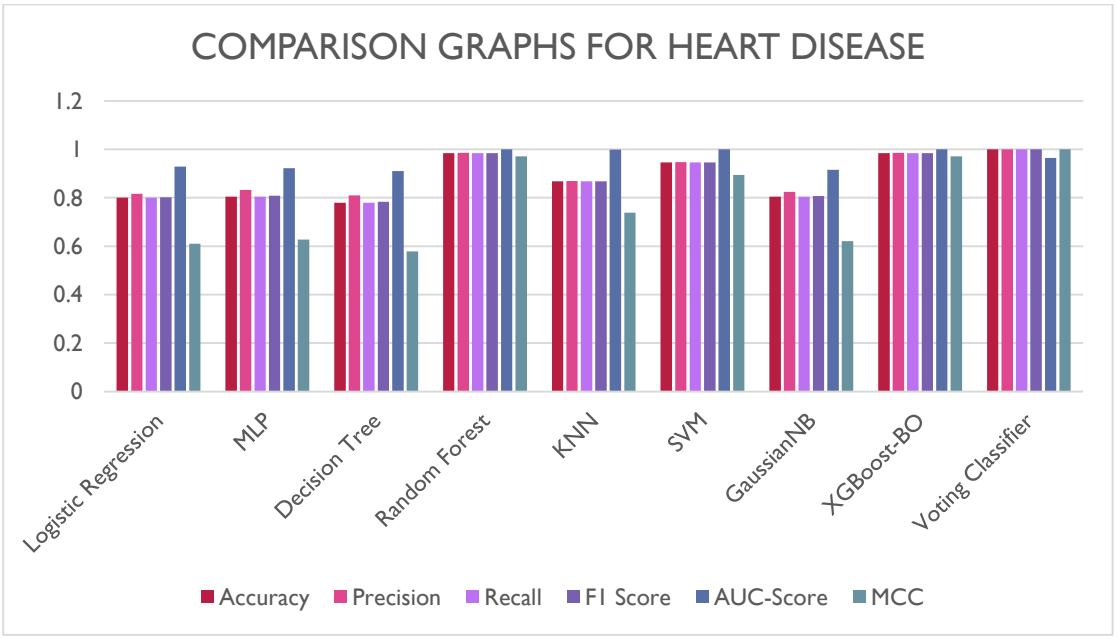
"Graph.4 Comparison Graphs for Erbil Heart Disease"



"Table.5 Performance Evaluation Metrics for Heart Disease"

ML Model	Accuracy	Precision	Recall	F1 Score	AUC-Score	MCC
Logistic Regression	0.800	0.817	0.800	0.802	0.929	0.610
MLP	0.805	0.833	0.805	0.808	0.922	0.627
Decision Tree	0.780	0.810	0.780	0.784	0.911	0.578
Random Forest	0.985	0.986	0.985	0.985	1.000	0.971
KNN	0.868	0.869	0.868	0.868	0.999	0.738
SVM	0.946	0.948	0.946	0.946	1.000	0.894
GaussianNB	0.805	0.824	0.805	0.807	0.916	0.621
XGBoost-BO	0.985	0.986	0.985	0.985	1.000	0.971
Voting Classifier	1.000	1.000	1.000	1.000	0.965	1.000

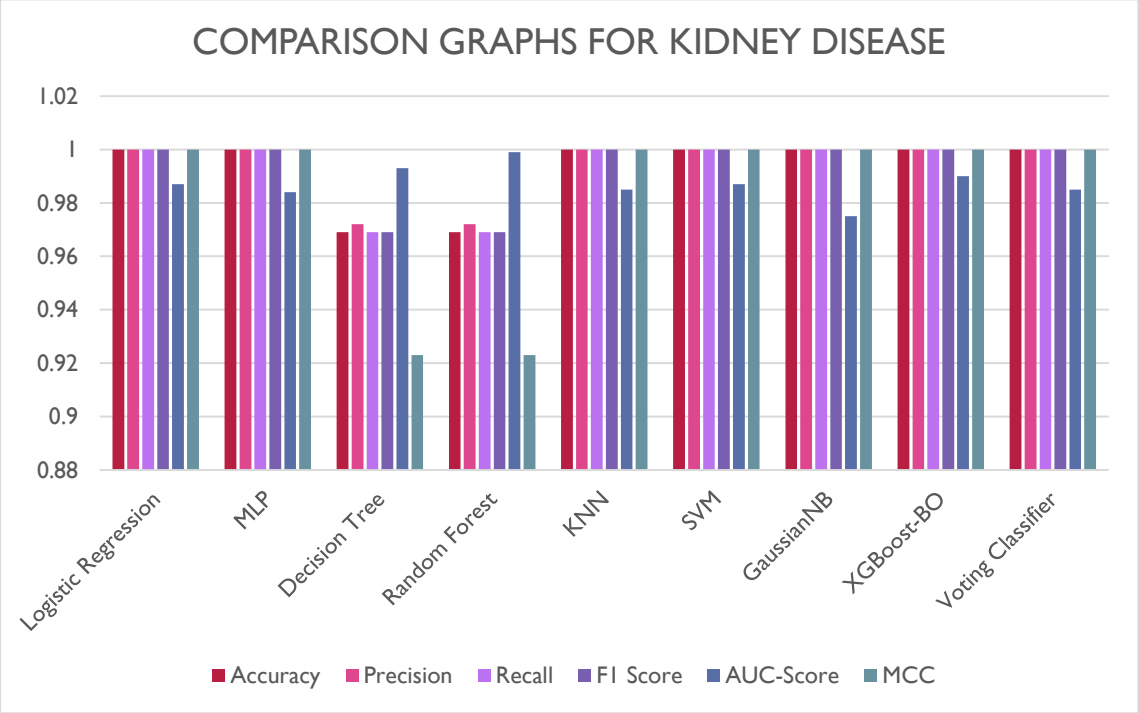
"Graph.5 Comparison Graphs for Heart Disease"



"Table.6 Performance Evaluation Metrics for Kidney Disease"

ML Model	Accuracy	Precision	Recall	F1 Score	AUC-Score	MCC
Logistic Regression	1.000	1.000	1.000	1.000	0.987	1.000
MLP	1.000	1.000	1.000	1.000	0.984	1.000
Decision Tree	0.969	0.972	0.969	0.969	0.993	0.923
Random Forest	0.969	0.972	0.969	0.969	0.999	0.923
KNN	1.000	1.000	1.000	1.000	0.985	1.000
SVM	1.000	1.000	1.000	1.000	0.987	1.000
GaussianNB	1.000	1.000	1.000	1.000	0.975	1.000
XGBoost-BO	1.000	1.000	1.000	1.000	0.990	1.000
Voting Classifier	1.000	1.000	1.000	1.000	0.985	1.000

"Graph.6 Comparison Graphs for Kidney Disease"

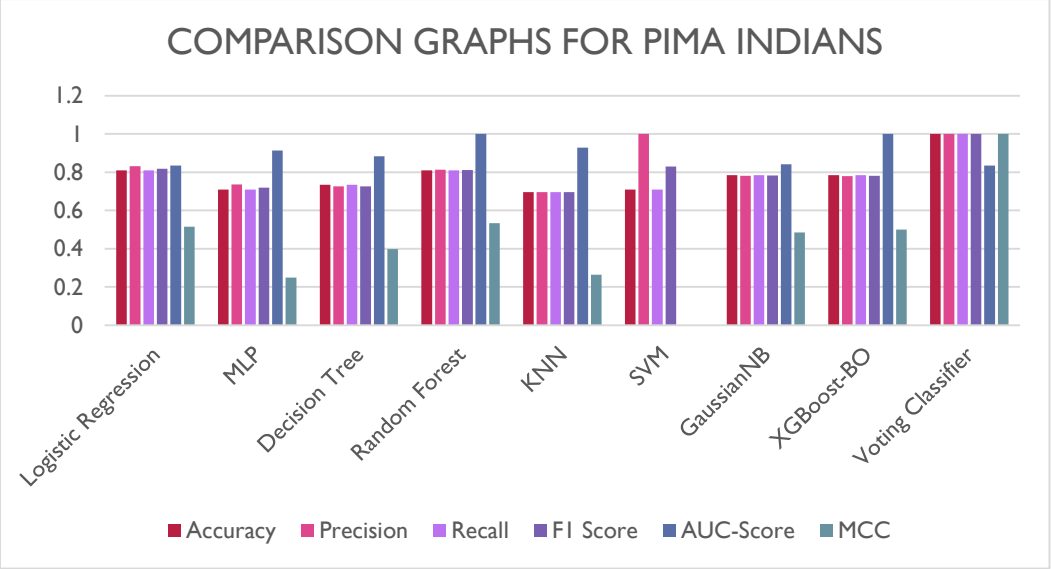


"Table.7 Performance Evaluation Metrics for Pima Indians"

ML Model	Accuracy	Precision	Recall	F1 Score	AUC-Score	MCC
Logistic Regression	0.810	0.832	0.810	0.818	0.835	0.515
MLP	0.709	0.737	0.709	0.720	0.913	0.250
Decision Tree	0.734	0.727	0.734	0.727	0.884	0.399
Random Forest	0.810	0.813	0.810	0.811	1.000	0.534
KNN	0.696	0.696	0.696	0.696	0.928	0.264
SVM	0.709	1.000	0.709	0.830	0.000	0.000

GaussianNB	0.785	0.782	0.785	0.783	0.842	0.485
XGBoost-BO	0.785	0.780	0.785	0.781	1.000	0.500
Voting Classifier	1.000	1.000	1.000	1.000	0.835	1.000

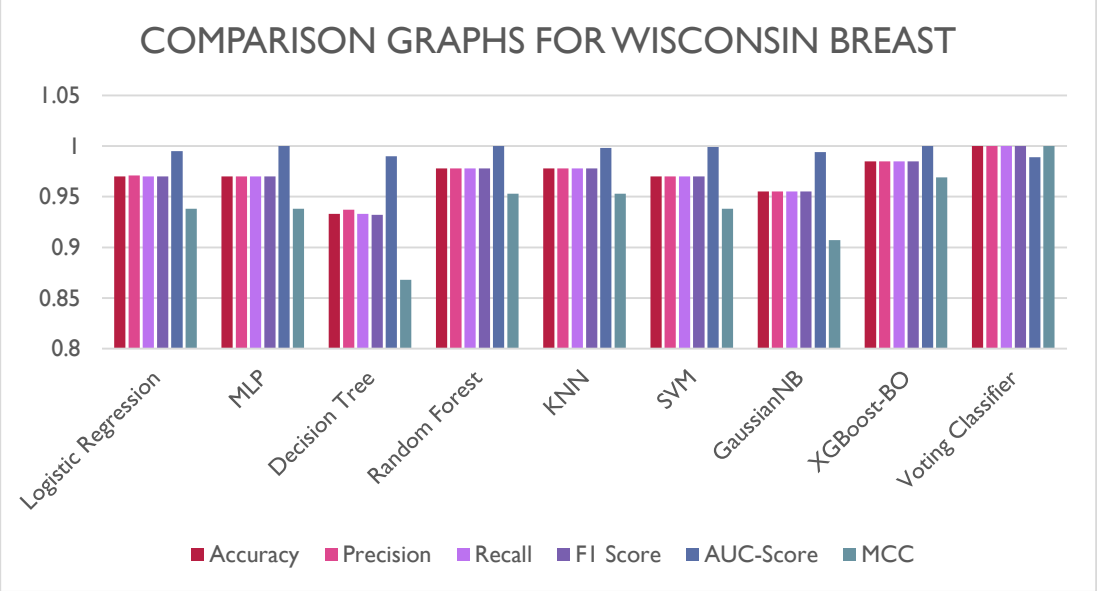
"Graph.7 Comparison Graphs for Pima Indians"



"Table.8 Performance Evaluation Metrics for Wisconsin Breast"

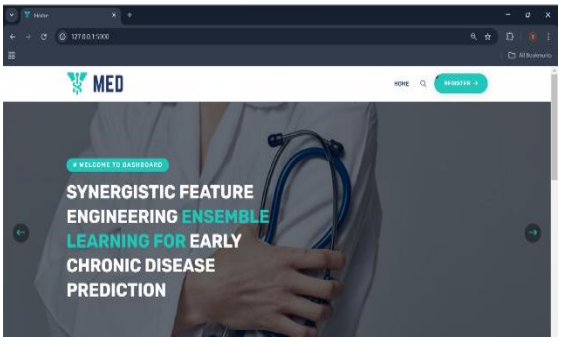
ML Model	Accuracy	Precision	Recall	F1 Score	AUC-Score	MCC
Logistic Regression	0.970	0.971	0.970	0.970	0.995	0.938
MLP	0.970	0.970	0.970	0.970	1.000	0.938
Decision Tree	0.933	0.937	0.933	0.932	0.990	0.868
Random Forest	0.978	0.978	0.978	0.978	1.000	0.953
KNN	0.978	0.978	0.978	0.978	0.998	0.953
SVM	0.970	0.970	0.970	0.970	0.999	0.938
GaussianNB	0.955	0.955	0.955	0.955	0.994	0.907
XGBoost-BO	0.985	0.985	0.985	0.985	1.000	0.969
Voting Classifier	1.000	1.000	1.000	1.000	0.989	1.000

"Graph.8 Comparison Graphs for Wisconsin Breast"



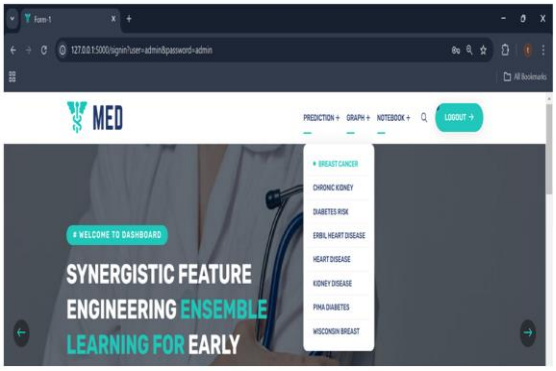
Graphs (1–8) display accuracy in mild blue, precision in orange, recall in gray, and F1-score in mild yellow; AUC-rating in blue and MCC in green. The balloting Classifier achieves the best values among the different models by showing better performance over all standards. The graphs above aesthetically show those results.

User Dashboard



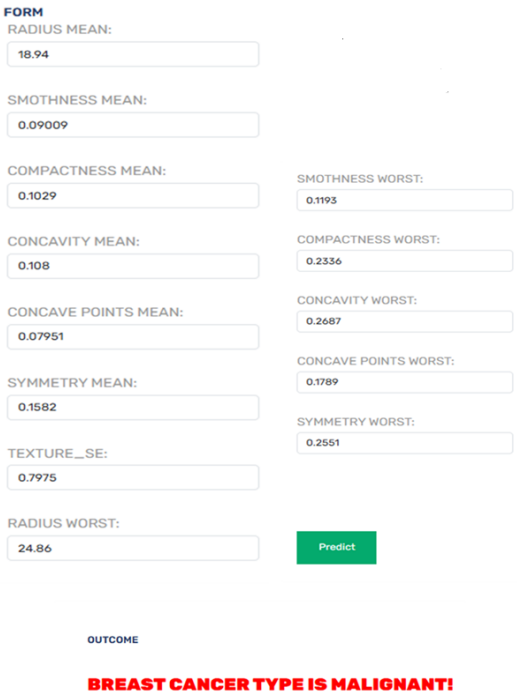
"Fig.10 Home Page"

this is a user interface dashboard found in figure 10 above; it serves as a welcome message for navigating pages.



"Fig.11 User home Page"

this is a user home page shown above figure 11; using this consumer will enable one to select type of class for prediction.



"Fig.12 Prediction result"

this is a result screen shown above figure 12; the user gets output for loaded input data.

5. CONCLUSION

Early diagnosis of chronic illnesses nonetheless provides a challenge for researchers, which fuels studies on several artificial intelligence methods for medical data evaluation and sickness prediction place to begin. using machine learning models has helped to significantly enhance prediction accuracy over a spectrum of persistent sicknesses. several algorithms were used to assess their predictive powers inside the study of datasets such as Breast cancer, continual Kidney, Diabetes danger, Erbil heart disorder, Kidney sickness, Pima Indians, Wisconsin Breast, and coronary heart sickness. amongst them, the voting Classifier—which aggregates several algorithms for stronger resilience—showcased always higher performance on all datasets. Its capability to mix the pleasant features of several fashions helped it to surpass different person classifiers, for this reason generating more accurate and dependable predictions. This makes it a useful instrument for early analysis because it gives scientific practitioners better decision-making energy. Such excessive-overall performance techniques allow one to greatly boom the prediction and diagnosis of chronic diseases, subsequently helping extra activate interventions and improved patient outcomes.

future developments could concentrate on using more diverse algorithms and tuning them for certain persistent diseases so further improving the performance of ensemble fashions such as the balloting Classifier. Early detection competencies may be enhanced by such as actual-time data from wearable fitness sensors and digital fitness records. furthermore helping to create greater openness and confidence in forecasts by using investigating deep learning strategies for function extraction and consisting of explainable artificial intelligence would help individualized healthcare subsequently.

REFERENCES

[1] W. H. O. Diet, ‘‘Chronic diseases,’’ World Health Org., Geneva, Switzerland, 2003.

[2] R. Sawhney, A. Malik, S. Sharma, and V. Narayan, ‘‘A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease,’’ Decis. Anal. J., vol. 6, Mar. 2023, Art. no. 100169.

[3] G. Viswanath "Machine-Learning-Based Cloud Intrusion Detection", in International Journal of Mechanical Engineering Research and Technology, 2024, Vol.16, pp 38-52.

[4] U. Ullah and B. Garcia-Zapirain, ‘‘Quantum machine learning revolution in healthcare: A systematic review of emerging perspectives and applica tions,’’ IEEE Access, vol. 12, pp. 11423–11450, 2024.

[5] G.Viswanath & Dr.G.Swapna "Diabetes Diagnosis Using Machine Learning with Cloud Security" in

Cuestiones de Fisioterapia, 2025, Vol.54, No.2, pp 417-431.

[6] X. Xiong, X. Guo, P. Zeng, R. Zou, and X. Wang, ‘‘A short-term wind power forecast method via XGBoost hyper-parameters optimization,’’ Frontiers Energy Res., vol. 10, May 2022, Art. no. 905155.

[7] X.-W. Chen and J. C. Jeong, ‘‘Enhanced recursive feature elimination,’’ in Proc. 6th Int. Conf. Mach. Learn. Appl. (ICMLA), Dec. 2007, pp. 429–435.

[8] G.Viswanath “Improved Light GBM Model Performance Analysis and Comparison for Coronary Heart Disease Prediction”, in International Journal of Information Technology and Computer Engineering, 2024, Vol.12, pp.658-672.

[9] A. Bohr and K. Memarzadeh, Artificial Intelligence in Healthcare. New York, NY, USA: Academic, 2020.

[10] G.Viswanath & Dr.G.Swapna “Health Prediction Using Machine Learning with Drive HQ Cloud Security” in Frontiersin Health Informatics, 2024. Vol. 13, No. 8, pp. 2755-2761.

[11] S. Agarwal and D. M. G. C. Prabha, ‘‘Chronic diseases prediction using machine learning—A review,’’ Ann. Rom. Soc. Cell Biol., vol. 25, no. 1, pp. 3495–3511, 2021.

[12] A. Maleki, M. Raahemi, and H. Nasiri, ‘‘Breast cancer diagnosis from histopathology images using deep neural network and XGBoost,’’ Biomed. Signal Process. Control, vol. 86, Sep. 2023, Art. no. 105152.

[13] M. J. Raihan, M. A.-M. Khan, S.-H. Kee, and A.-A. Nahid, ‘‘Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP,’’ Sci. Rep., vol. 13, no. 1, p. 6263, Apr. 2023.

[14] S. Hegde and M. R. Mundada, ‘‘Early prediction of chronic disease using an efficient machine learning algorithm through adaptive probabilistic divergence based feature selection approach,’’ Int. J. Pervasive Comput. Commun., vol. 17, no. 1, pp. 20–36, Feb. 2021.

[15] E. Maini, B. Venkateswarlu, D. Marwaha, and B. Maini, ‘‘Upgrading the performance of machine learning based chronic disease prediction systems using stacked generalization technique,’’ Int. J. Comput. Digit. Syst., vol. 10, no. 1, pp. 1031–1039, Nov. 2021.

[16] T. H. H.Aldhyani, A. S. Alshebami, and M. Y. Alzahrani, ‘‘Soft clustering for enhancing the diagnosis of chronic diseases over machine learning algorithms,’’ J. Healthcare Eng., vol. 2020, pp. 1–16, Mar. 2020.

[17] O. A. Jongbo, A. O. Adetunmbi, R. B. Ogunrinde, and B. Badeji-Ajisafe, ‘‘Development of an ensemble approach to chronic kidney disease diagno sis,’’ Sci. Afr., vol. 8, Jul. 2020, Art. no. e00456.

[18] E. Listiana, R. Muzayanah, M. A. Muslim, and E. Sugiharti, ‘‘Optimiza tion of support vector machine using information gain and AdaBoost to improve accuracy of chronic kidney disease diagnosis,’’ J. Soft Comput. Explor., vol. 4, no. 3, pp. 152–158, 2023.

[19] Breast Cancer Wisconsin (Diagnostic) Data Set. Accessed: Nov. 2023. [Online]. Available: cancer-wisconsin-data https://www.kaggle.com/datasets/uciml/breast

[20] L. O. D. Chicco and C. A. Lovejoy. (2021). Chronic Kidney Disease EHRs Abu Dhabi. [Online]. Available: https://www.kaggle.com/datasets/davidechicco/chronic-kidney-disease-ehrs-abu-dhabi

[21] Early Stage Diabetes Risk Prediction Dataset. Accessed: Nov. 2023. [Online]. Available: https://archive.ics.uci.edu/dataset/529/early

[22] Erbil Heart Disease Dataset. Accessed: Nov. 2023. [Online]. Available: https://www.kaggle.com/datasets/hangawqadir/erbil-heart-disease-dataset

[23] Heart Disease Dataset. Accessed: Nov. 2023. [Online]. Available: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset?select=heart.csv

[24] Kidney Disease Dataset. Accessed: Nov. 2023. [Online]. Available: https://www.kaggle.com/datasets/akshayksingh/kidney-disease-dataset

[25] Pima Indians Diabetes. Accessed: Nov. 2023. [Online]. Available: https://data.world/uci/pima-indians-diabetes

[26] G.Swapna & K Bhaskar “Malaria Diagnosis Using Double Hidden Layer Extreme Learning Machine Algorithm With Cnn Feature Extraction And Parasite Inflator” International Journal of Information Technology and Computer Engineering, Vol. 12 , 2024, pp536-547,DOI: https://ijitce.org/index.php/ijitce/article/view/704

[27] G.Swapna, “A Drug-Target Interaction Prediction Based on Supervised Probabilistic Classification” Journal of Computer Science, Vol.19, 2023, pp.1203-1211, DOI: https://doi.org/10.3844/jcssp.2023.1203.1211

[28] Viswanath Gudditi, “A Smart Recommendation System for Medicine using Intelligent NLP Techniques”, 2022 International Conference on Automation, Computing and Renewable Systems (ICACRS), 2022, pp.1081-1084.

[29] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, ‘‘Disease prediction by machine learning over big data from healthcare communities,’’ IEEE Access, vol. 5, pp. 8869–8879, 2017.

[30] G.Swapna & K Bhaskar “Early-Stage Autism Spectrum Disorder Detection Using Machine Learning “International Journal of HRM and Organizational Behavior, pp269-283, DOI: https://ijhrmob.org/index.php/ijhrmob/article/view/242

[31] Wisconsin Breast Cancer Database. Accessed: Nov.2023. [Online]. Avail able: https://www.kaggle.com/datasets/roustekbio/breast-cancer-csv

[32] Viswanath “Multiple Cancer Types Classified Using CTMRI Images Based On Learning Without Forgetting Powered Deep Learning Models”, in International Journal of HRM and Organizational Behavior, 2024, Vol.12, pp 243-253.

[33] S. Palaniappan and R. Awang, ‘‘Intelligent heart disease prediction system using data miningtechniques,’’ in Proc. IEEE/ACSInt. Conf. Comput.Syst. Appl., Apr. 2008, pp. 108–115.

[34] P. Hamet and J. Tremblay, ‘‘Artificial intelligence in medicine,’’ Metabolism, vol. 69, pp. S36–S40, Jan. 2017.

[35] O. Khan, J. H. Badhiwala, G. Grasso, and M. G. Fehlings, ‘‘Use of machine learning and artificial intelligence to drive personalized medicine approaches for spine care,’’ World Neurosurgery, vol. 140, pp. 512–518, Aug. 2020.