International Journal of Advanced Computer Technology
ISSN: 2319-7900 www.ijact.org Volume 13, Issue 1, February 2024

Cross Validation Machine Learning Model Predicts More Accurate: A Comparative Study of Heart Disease Using Linear Regression, Support Vector Machine, K Neighbors and Random Forest Models

Yagyanath Rimal[1], Siddhartha Paudel [2], Abeer Alsadoon[3,4,5], Madhav Prasad Koirala[1], Sumeet Gill[6]

[1] Pokhara University, Nepal(rimal.yagya@gmail.com)

[2] IOE, Pulchowk Campus, Patan, Nepal(paudelsiddhartha36@gmail.com)

[3]School of Computing Mathematics and Engineering, Charles Sturt University (CSU), School of Computer Data and Mathematical Sciences, Australia

[4]Western Sydney University (WSU), Sydney, Australia

[5]Asia Pacific International College (APIC), Sydney, Australia(alsadoon.abeer@gmail.com)

[1]Pokhara University, Nepal (mploirala@pu.edu.np)

[6]Maharshi Dayanand University Rohtak(drsumeetgill@mdrohtak.ac.in)

Correspondence Author: Yagyanath Rimal, Pokhara University, Nepal, rimal.yagya@gmail.com

Abstract: This primary research paper focuses on using cross-validation, where each iteration of test data is uniquely structured to ensure optimal model performance by combining weak learners for improved model final accuracy. In the machine learning process, data is commonly split into two sets: a training set comprising 70% of the data and a test set comprising 30%. Cross-validation is then utilized for training and evaluation, often involving reusing previous data sets. This research study transforms the original datasets and cross-validating comparative analysis using LR, SVM, KNN, and RF methodologies to predict heart disease. The objective is to easily identify the average accuracy of model predictions and subsequently make recommendations for model selection based on both cross-validated increased (5 to 13%) and non-cross-validated approaches. From comparing each model's accuracy scores, it is found that the logistic regression and k-nearest neighbour models achieved the highest accuracy of 81% among the four models.

Similarly, the random forest model attained an F1 score of 95%, indicating the highest accuracy score from the enhanced heart disease sample. These findings can be further corroborated using learning curve validation.

Conversely, the linear regression model exhibited the lowest accuracy of 84% among the four machine learning models.

Keywords- Machine Learning, Cross-validation, Accuracy-precision, Learning Curve, Health informatics, Bio-signal processing

1. Introduction

Machine learning involves crafting models based on training datasets, which are evaluated using testing datasets of unseen samples. While the train-test split is a common practice for dividing research datasets into training and testing sets, it is often less preferable for model prediction. Another option is splitting available data (training/testing) sets before with some ratio 70:30 splits where the programmer builds the model using training data and then whose value is further tested with test unseen data sets. This approach achieves greater accuracy than the initial option, but it might not be suitable if a student asks questions beyond the chapters taught to attain the highest grade. Cross-validation is a method of training a model by storing some portion of the sample data set of each split and the rest of the data set to train the model (Maldonado et al.). Similarly, the authors(Mahesh et al.) examined stratified cross-validation employed to split the data, ensuring a similar distribution of target outputs among prediction samples, thereby yielding the best average score. The holdout method functions by reserving a portion of the training dataset for model validation. In contrast, stratified n-fold cross-validation effectively handles imbalanced datasets, ensuring each fold contains a proportional representation of each output class. Leave-p-out cross-validation entails training with n/p samples and validating with p points, repeating this process for all combinations, and averaging the errors until randomness is minimized (Schmidt et al.). Linear regression, random frost support vector machine, bootstrapping, and cross-validation techniques are common algorithms used to solve overfitting problems in medical research. Bootstrapping uses the remaining sample data to resample the data, while cross-validation techniques use large features to compare disease responses (Ye et al.). The author (Gimenez-Nadal et al.) split the dataset into nfold, trained the model on the training set, and validated it on the test set. Repeat these steps 3 to 6,000 times, with the first convolution reserved for model testing and the rest used for model training. Bias measures the difference between the model's prediction and the target value, while variance measures the disagreement of different predictions across different datasets. In an ideal scenario, the model strikes an optimal balance between bias and variance(Dodge et al.). The methodology used for data splitting significantly impacts the accuracy of model prediction. Similarly, the authors (Belkin et al.) employed the term 'generalization' to describe the effectiveness of a model in extracting useful data patterns and accurately classifying unseen data samples. Overfitting models remember the data patterns of the training dataset but do not generalize to unseen data, leading to high model variance(Kernbach and Staartjes). Underfitting arises when the model fails to extract patterns from the dataset adequately, often due to insufficient or noisy training data. The objective is to achieve an optimal fit that accurately captures patterns within the training data. Similarly, the author (Olaniyi et al.) proposes a three-phase demo based on artificial neural networks. The angina analysis model showed a classification accuracy of 88.89% using the UCI dataset. The neural network backpropagation model showed 85% accuracy when testing. Similarly(Benjamin et al.), the authors explored inquiries regarding nonsmoking among children aged 12 to 19 years, which increased from 76% to 94%. However, factors such as physical activity, body mass index, and blood glucose levels did not show improvement; instead, the prevalence declined from 70% to 60% over the same period(Arora et al.). The classification framework and accomplished framework show 89.1 % accuracy; however, model wise differs by 80.09% to 95.91% individually utilizing ventricular systolic execution within the distribution the distributed reports shift broadly from 13% to 74%; the detailed yearly mortality rate moreover shifts from 1.3% to 17.5%. Similarly, the authors (Zuhair et al.) combined medical decision-making with a framework for cardiac infection symptoms using machine learning classifiers such as multilayer perceptions and artificial neural networks. Their methodology uses machine learning algorithms and analytical hierarchical fuzzy processing within artificial neural networks to diagnose heart disease. Their proposed classification system achieved a classification accuracy of 91.10%. This work mainly discusses the model selection and accuracy without dealing with various cases of overfitting and underfitting classification computation double classification problems, y [0, 1], negative history, and an estimate of the forward variable y for one positive course. The multiclass to predict estimates of y for y [0, 1, 2, 3]. A guess is sketched to classify two classes and 1 class. The yield of the classifier is 0.5. A support vector machine can be a machine learning classification computation commonly used for classification problems. The support vector machine used the most extreme edge technology modified (Xiong et al.). According to (Nadar and Kamatchi) imbalanced datasets were employed to classify primary school and higher education using online multiple-choice tests from Bharathiar University. The research revealed that approximately 20% of students in the USA do not complete their graduation on time, while in Europe, the range is between 20% to 50%. Likewise, the authors (Khan and Ghosh) examined relevant studies published between 2000 and 2018, revealing that multiple factors influence performance in non-linear ways within online learning contexts. The analysis focused on identifying influencing factors based on assessment behaviours, association rule mining, and regression and classification analyses for performance prediction. A significant majority, accounting for 46% of modelling studies, preferred to classify performance as either success or failure. Similarly, the author (Yousafzai et al.) used supervised mastering algorithms to improve a predictive version of Federal Board of Intermediate and Secondary Schooling Islamabad Pakistan, using ok = 10. In k-fold pass-validation, a reduced education vector-based totally aid vector device is proposed to predict at-risk and marginal college students whose support vector completed a training vector discount of at least fifty-nine.7% without changing the margin or accuracy of the classifier. Moreover, the effects confirmed the proposed approach achieved a basic accuracy of 92.20–93.8% and 91. Three 93.5% in predicting at-risk, respectively. Likewise, in their study referenced as (Smirani et al.), the authors employed light gradient boosting, extended gradient boosting, random forest, and multilayer perceptron classifiers sourced from UCI records. They categorized the data into three groups for error prediction. Additionally, they explored stack generalization in machine learning repositories. Their results showcased impressive metrics, including an average sensitivity of 97.3%, joint accuracy of 97.2% in classification, an F1 rating of 97.1%, and an average of 98.86% for the neural network algorithm. Moreover, they observed a significant decrease in the dropout rate, from 12% to 1.14%.Similarly, the author (Usama et al.) used a neural community version to exhibit that the proposed model reaches up to an accuracy of ninety-five 71%, higher than many present methods for cerebral infarction disorder. Likewise, in the study by the authors referenced (Shukla et al.), DBSCAN was utilized to identify and extract nine clusters of informative gene data, selected via differential gene expression analysis, which were then classified into five distinct categories. Subsequently, a deep learning approach was employed to ensemble the outputs of these five classifiers. Similarly, the authors state that the modified J48 classifier is used to boost the accuracy fee of the data mining technique. MATLAB's facts mining tool generates the WEKA's decision classifiers and Naive Bayesian classifiers. the general accuracy is around eighty-three. Likewise, a memetic algorithm was utilised in the study by the authors referenced (Naz and Ahuja), improving accuracy from 88.0% to 93.2%. Additionally, it was observed that the memetic algorithm outperformed both the genetic algorithm and a regression model in terms of accuracy. Similarly (Dharma et al.) uses a genetic algorithm-primarily based regression model for predicting inflation levels. The version becomes educated and evaluates the usage of facts. Similarly the authors (Touzani et al.), (Mohan et al.) used a prediction model delivered with one-of-a-kind combos of features and several acknowledged category strategies. We produce an improved performance level with an accuracy degree of 88.7% via the prediction version for coronary heart disorder with the hybrid random forest area with a linear model. Similarly authors (Anuradha and Velmurugan), the prediction of performance was conducted using the k-nearest neighbor algorithm. The overall accuracy of the tested classifiers reached 60%. Notably, the decision tree achieved an accuracy of approximately 72.51% in 10-fold cross-validation testing and 69.66% in split testing. Precision was notably high for the primary class (67-76%) and the second class (72-85%). Likewise, authors referenced as (Hussain and Dimililer) conducted research to identify the most influential feature of the target class and to determine which method surpasses the commonly used RF, Component, J48, and Bayes classifiers. By incorporating socio-economic, demographic, and educational data, the random forest model achieved an impressive accuracy of 99%. They also analyzed the internal factors affecting the final semester percentage. Similarly, author (Smirani et al.) used the demographics to outperform random forest by providing 99.90% accuracy on training information 10-fold cross-validation and 99.82% accuracy on the holdout method. When implementing guidance, the accuracy of the basic ANN may increase by up to 100% during training, although the accuracy during testing/validation may vary significantly for prediction. Conversely, other methods may exhibit the opposite effect. Self-efficacy and motivation for success are particularly relevant when addressing heart disease and its significant factors, both post-diagnosis and during pre-symptomatic stages, aiming for reduction. Machine learning plays a giant position in extracting the hidden capabilities from the scientific records beneficial for early detection from the heart ailment report repository, that's the reason approximately 12 million dying happens in globally (Chowdhury et al.). Coronary disorder dying is observed greater in the USA than in other advanced Europe (Townsend et al.). Hence, based on the literature discussed above, the selection of machine learning models and the method of splitting data samples before model training significantly impact the accuracy scores for classification and prediction tasks. This research tries to expect the correct machine mastering version validation accuracy prediction using four linear regressions, aid vector device nearest neighbour, and random woodland version validation of contracts in the coronary arteries.

2. Data Preparation Flow Chart

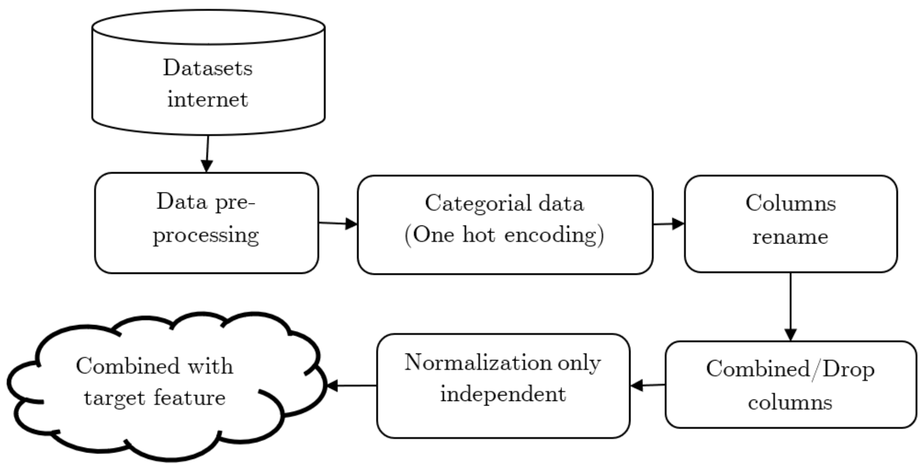

Figure 1. Data preparation flow diagram

The heart disease data sets contain 13 categorical attributes whose data needs to be preprocessed before machine learning model testing and evaluation; after loading the dataset into the Python console, the Python command df. Types describe the data types categories with their respective categories. Every unique feature accompanied by a specific 'c' command delineates the characteristics of the datasets. When utilizing the Onehotencoder with the parameter 'categories='auto'', it is fitted using the fit transform() method on the categorical independent features such as 'sex', 'cp', 'FBS', 'resting', 'exacting', 'slope', 'thal', and 'ca', where numbers represent different categorical values using to array function. The column name of each categorical Onehotencoder constitutes 76 columns of data sets of each feature. Cross-validation provides better model optimization of heart disease using linear regression and support vector machine-learning model before finalizing the best model for the research dataset. The most important details in this text are the age, chest pain, trest-bps, chol, thalach, old peak, m, f, typical angina, atypical angina, non-anginal pain, asymptomatic, normal, abnormal, normal, abnormal, yes, no, upsloping, flat, down, normal, fixed defect, reversible defect, non, Ca0, Ca1, ca2, ca3, ca4's (Ansari et al.).

ISSN: 2319-7900 www.ijact.org Volume 13, Issue 1, February 2024

Likewise, in the study by the author referenced (Amarbayasgalan et al.), the heart disease dataset consisted of samples with 14 independent variables and a final target variable indicating the presence (1) or absence (0) of heart disease. This dataset comprised 303 records of heart disease patients, with 381647 views and 62705 downloads as of January 2024. After loading the SKlearn preprocessing library of standard scaler into the python console, rename their respective columns as final2['age', 'trestbps', 'chol', 'thalach', 'old peak', m,f, 'normal' (Barhoom et al.).

Table 1. Data table preparation before and after

age	trestbps	chol	thalach	old peak		age	trestbps	chol	thalach	Oldpeak
63	145	233	150	2.3	Before After	.952	.763	-0.256	.0154	1.08
37	130	250	187	3.5		-1.91	-.092	.0721	1.633	2.122
41	130	204	172	1.4		-1.47	-0.092	-.816	.977	.3109
56	120	236	178	0.8		0.18	-0.663	-.198	1.239	-0.206

After combining the target column with the normalized data set, it becomes complete for four model comparisons. This research forest used four model comparisons using normalized with 80: 20 splits, and the parameters using stratify in each of these datasets, the target/label data proportion is preserved as 50:50 when for the classes [0,1]. It indicates that there would not be an oversample or under-sampling problem in both training and test sets. Setting the random_state to 42 ensures identical training and test sets across various runs. However, when random_state is set to 0, the training and test sets differ from the previous case. This discrepancy directly impacts the performance score of the model.

3. Validation Design Diagram

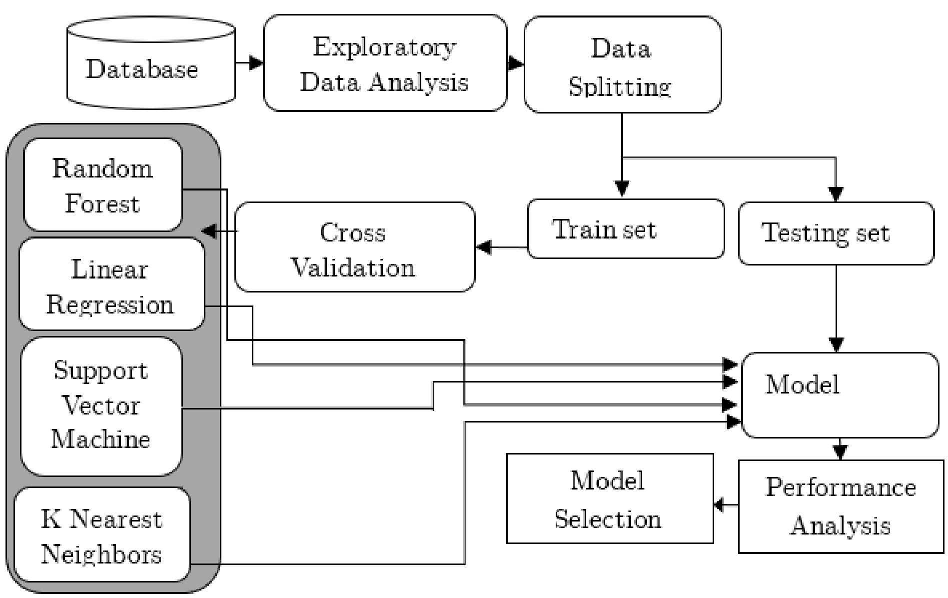

Figure 2. Model analysis process

This research tries to validate the best machine learning model using four algorithms, whose procedures execute individual model accuracy prediction and then cross-validation procedure with 5-fold data splits of tested after-train test splits of machine learning heart disease preprocessing data sets. This output was plotted further using a learning curve with 10-fold cross-validation. So, data scientists received the best model.

4. Results and Discussions

After designing the data sets, the correlation between dependent and independent variables is described using heatmap (final2.corr(), cmap='cool warm') in the case for displaying each correlation value "SNS. heatmap (final2.corr(), annot=True)". Correlation plots are used to understand which variables are related to each other and the strength of this relationship.

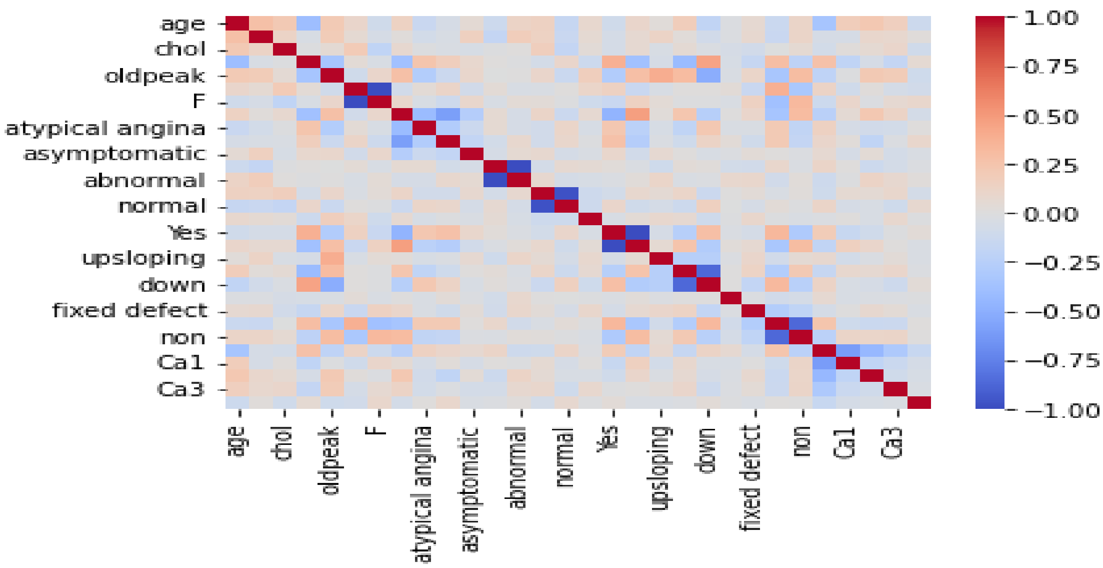

Figure 3. Data correlation heat map

From the above Figure 2, the heart disease dataset, the correlation coefficients between dependent and independent features, as determined by logistic regression, reveal a mean squared error of 0.18, signifying satisfactory performance. The coefficient of determination, at 0.27, suggests that the model explains 27% of the variability in the data. This statistical analysis helps evaluate the model's ability to explain and predict future outcomes. Additionally, a value of 0.30 for the coefficient of determination indicates that 9% of the variance between the variables is shared or common. Similarly, when applying the support vector machine, the mean squared errors greater than logistic regression is 0.20, and the coefficient of determination decreased is 0.21, respectively. The output 0.21 suggests that the independent predicts 21% of the dependent variable. From the logistic summary table provided, the R-squared measure indicates that the independent variables explain 58% of the variance in the dependent variable. However, the adjusted R-squared, a more accurate measure, is 54%. The p-value (P), close to zero, suggests strong evidence against the null hypothesis. The F-statistic (F), utilized to test the overall significance of the regression model, is 17.71. A higher F-statistic signifies a more substantial relationship between the independent and dependent variables.

The intercept represents the expected value of the dependent variable when all independent variables are zero. In this case, the intercept is 0.54, with a high t-value and low p-value. The coefficients indicate that a one-unit increase in Age is associated with a 0.20-unit increase in the dependent variable, but it is not statistically significant. The coefficients indicate that a one-unit increase in Trestbps is linked with a 0.41-unit decrease in the dependent variable. Notably, the Sex_1 variable demonstrates statistical significance, with a t-value of 2.64 and a low p-value (0.001). However, the other variables (Age, Trestbps, Chol, Thalach, Oldpeak) do not exhibit a significant relationship with the dependent variable. Hence, further research is warranted to assess the prediction accuracy of heart disease patients using machine learning models with cross-validation.

Table 2. Logistic summary statistics

R-squared:0.58	Adjusted R-squared:0.54		P: 5.38e-41	F-stasticts:17.71
	coef	std	t	P
Const	0.54	0.019	28.29	0.00
Age	0.20	0.024	1.023	0.30
Trestbps	-0.41	0.021	-0.92	0.56
Chol	-0.016	0.026	1.67	0.42
Thalach	0.042	0.027	-1.80	0.9
Oldpeak	-0.041	0.022	-3.40	0.07
Sex_1	-0.07	0.020	2.64	0.001
Fbs_1	0.011	0.043	0.58	0.00

5. Machine Learning Model Without Using CV

The default machine learning model of four different machine learning model parameters (models = [LogisticRegression(max_iter=1000), SVC (kernel='linear’), KNeighborsClassifier (), RandomForestClassifier ()]) were designed in the model and then using loop the model whose accuracy score was calculated using after fitting the models. This process needs to split data into test and train splits model design, predict with test data, and calculate each model accuracy score. The machine learning model fit (train, train), test_data_prediction = model. Predict (test), accuracy = accuracy score (test, test_data_prediction), print (‘Accuracy score of the ‘, model,’ = ‘, accuracy). The console output reveals that the logistic regression and k-nearest neighbours models achieved the highest accuracy at 81.9%, followed by the support vector machine and random forest models with an accuracy of 78% each. Specifically, the accuracy scores are as follows: LogisticRegression (81.9%), SVC with linear kernel (78.6%), KNeighborsClassifier (81.9%), and RandomForestClassifier (78.6%). These accuracy scores depict the models’ performance using default parameters. Furthermore, the accuracy classification score for multilevel problems indicates the exact extent to which the true labels in the y_train sample match. The confusion map generated for each class version outputs a plot_confusion_matrix(version, X_train, Y_train), and the confusion matrix plots, as shown in Figure 4, are utilized to evaluate class-specific errors in the model. The rows represent the actual classes of outcomes, while the columns represent the predictions made by the model. This table makes it easy to peer which predictions are wrong. The above model's accuracy scored using classification report is under train test splits of whole x and y, which is based on large variation; therefore, we need to cross-validate, which is selected after 5 integrations each time the test sample differs—the print (classification report (Y, model. predict(X))).

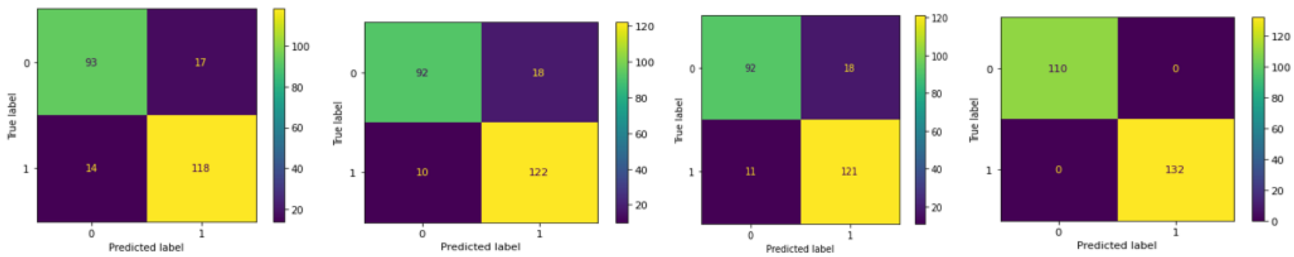

Figure 4. Confusion Matrix of Each Model

Table 3. When Maximum Accuracy

	Accuracy %	Precision	Recall	F1-score
LogisticRegression	81.9	0.86	0.88	0.87
SVC	78.6	0.88	0.91	0.88
KNeighborsClassifier	81.9	0.89	0.92	0.89
RandomForestClassifier	78.6	0.96	0.97	0.96

Table 4: When Minimum Accuracy

	Accuracy %	Precision	Recall	F1-score
LogisticRegression	81.9	0.86	0.83	0.84
SVC	78.6	0.85	0.81	0.85
KNeighborsClassifier	81.9	0.86	0.83	0.86
RandomForestClassifier	78.6	0.95	0.93	0.95

From the above table, the logistic regression and k nearest neighbour's algorithm predict a better result than the support vector machine and random forest model. When evaluating precision and F1 scores, the random forest model demonstrates the highest prediction accuracy at 97%. However, the precision of the random forest classifier is 96%, indicating the highest recall score. Nevertheless, the researcher recommends opting for the model with the lowest accuracy of the available models.

6. Default Machine Learning Model with Using CV

Another method of resampling heart disease datasets for machine learning is through cross-validation (CV). CV involves evaluating multiple K-fold models by training each on subsets of the data. The final prediction is determined by aggregating the results from evaluating these models on complementary subsets of the data. This approach is effective for detecting overfitting issues and promoting the generalization of patterns during model design. Each model individually evaluates and fitted and calculated their accuracy score using: cv_score_lr = cross_val_score (LogisticRegression(max_iter=1000), X, Y, cv=5), print(cv_score_lr), mean_accuracy_lr = sum(cv_score_lr)/len(cv_score_lr), mean_accuracy_lr = mean_accuracy_lr*100, mean_accuracy_lr = round (mean_accuracy_lr, 2), print(mean_accuracy_lr). Based on the table provided, the average accuracy score of the random forest and k-nearest neighbour models was the highest, achieving 84.15%. Linear regression followed closely behind with an average accuracy score of 83.81%, while the support vector machine model attained an accuracy of 82.49% after individual default model and averaging.

7. Machine Learning Model Using CV (Combined)

Similarly, the machine learning mode after using loop using five-fold cross-validation executes models = [LogisticRegression(max_iter=1000), SVC(kernel='linear'), KNeighborsClassifier(), RandomForestClassifier()] for comparing models cross-validation () for the model in models cv_score = cross_val_score (model, x,y, cv = 5), mean_accuracy = sum(cv_score) /len(cv_score), mean_accuracy = mean_accuracy*100,mean_accuracy = round (mean_accuracy, 2) produces the following table accuracy of each folds sample data. From the model accuracy score, the k nearest model produces the highest accuracy when (84.15%) and the second lowest accuracy from both the model Linear regression and Random Forest (83.83%). The lowest model accuracy from the Support vector machine scored (82.2%).

Therefore, a researcher might take the highest or lowest scores to evaluate the model accuracy for heart disease. The model might be confused because it takes max/min from the five cross-validated accuracies. The accuracy score using the Max/ Min of each model return value depends on the setting for the normalized parameter due to sample reshuffled using stratified value become true when researcher considered sample reshuffled when cross-validation iteration the difference between each model matters large model accuracy for correctly classified samples. Based on the bar plot above, the red bars represent the accuracy scores of the machine learning models, while the blue bars indicate the accuracy of each model with cross-validation. Notably, the linear regression model defaulted to 78% accuracy, but with cross-validation, it improved. The support vector machine model also exhibited a significant difference, increasing from 82% without cross-validation to a higher score with cross-validation. Interestingly, the k-nearest neighbour's model yielded the best results among the four models considered. Similarly, comparing a single independent model vs multiple with CV model support vector differs by 15% compared to k. The nearest model accuracy is below 5%.

Table 5. Accuracy scores of machine learning models using cross-validation

Model/Iteration	1	2	3	4	5	Average
Linear Regression	0.88	0.88	0.80	0.83	0.78	83.81
Support Vector	0.88	0.88	0.75	0.81	0.78	82.49
K-Nearest Neighbor	0.85	0.86	0.81	0.85	0.81	84.15

Random Forest	0.83	0.90	0.80	0.85	0.81	84.15

Table 6. Machine learning model accuracy using combined CV

Model/Iteration	1	2	3	4	5	Average
Linear Regression	88.5	0.88	0.80	0.83	0.78	83.81
Support Vector	88.5	0.88	0.75	0.81	0.78	82.49
K-Nearest Neighbor	85.2	0.86	0.81	0.85	0.81	84.15
Random Forest	86.8	0.86	0.78	0.86	0.8	83.83

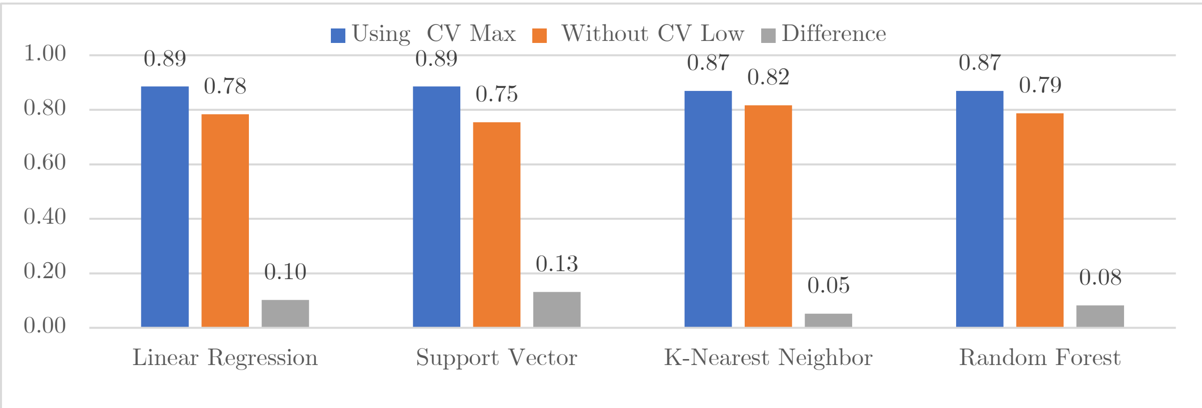

Figure 5. Min-Max bar chart accuracy comparison

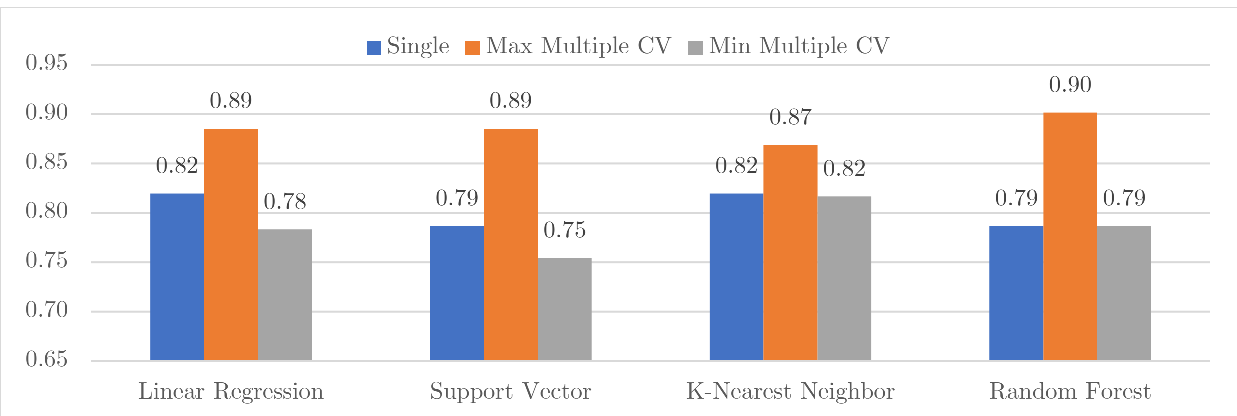

Figure 6. Represent Bar chart comparison of ensemble models

In Figure 6, the logistic regression default model achieved an accuracy of 82%, whereas with cross-validation, it improved to 89%. Similarly, the support vector machine model scored 79% by default, but its accuracy increased to 89% with cross-validation. The k-nearest neighbour model attained an accuracy of 82% by default, whereas with cross-validation, it achieved 87%. Lastly, the random forest model obtained 79% accuracy by default, but its accuracy soared to 90% with cross-validation. Similarly, it is concluded that the individual model has the least accuracy with mean values. Therefore, it is recommended that the max-multiple cross-validation model produce the highest accuracy. The random forest model with CV scored 90% accuracy compared to 75%.

8. Learning Curve of All Models

The learning curve represents the price of getting to know while extending through the years or repeated experiences. gaining knowledge of curves is a visualization of the difficulty predicted in learning a subject over time in addition to relative progress for the duration of the manner of getting to know. The concept is founded on a doubling of output, where a 70% learning curve indicates that the cumulative average time taken per unit decreases to 70% of the previous cumulative average time as the output doubles. The cumulative average time per unit is calculated from the initial unit-produced estimator while executing models or functions. GridSearchCV employs an absolute number of training examples to generate the curve. The scoring metric is utilized to evaluate the performance of the model version and determine optimal hyperparameters. If unspecified, the default metric is an estimator's score, set to five. However, in this study, the researcher opted for 10-fold cross-validation. The n_jobs symbolizes the wide variety of jobs to be run in parallel, and -1 indicates the application of all processors. After importing the learning curve package in the Python console, the normalized data first splits into dependent a and independent set of heart disease X=final4.drop([‘non’, ‘target’], axis=1) and y=final4 with the target. The learning rate splits with scoring accuracy, and the learning rate starts from 0.01, 1, 50, and 100 iteration splits. After the train test splits, the means of accuracy of the K nearest model plot is calculated. The learning curve describes the training and validation metric for overfitting and underfitting.

The above line indicates the validation curve changes gradually, and the lower line indicates the training error/accuracy score. This curve illustrates the evolution of error metrics as the model progresses in training and validation. Each line represents the collective impact of the model on heart datasets. Initially, the steep training line indicates rapid learning as the model reaches up to 150 training sets. Subsequently, both lines gradually decrease as the model improves its performance. However, beyond 250 iterations, the output becomes relatively constant, suggesting that the heart disease datasets exhibit high variance. Similarly, when the learning curve for Random Forest was generated after 150 training samples, the machine learning model exhibited a sharp increase in accuracy score but also indicated high bias compared to the dotted line.

The support vector machine and logistic regression studying the use of heart sickness data set a step-by-step process to validate that after 50 generations, the training facts curve is greater hastily than the validation curve, which, in the end, suggests overfitting. The curve serves various purposes, including evaluating different algorithms, selecting model parameters during design, and determining the data used for training. This variance in the relationship between practice and proficiency over time is called the ‘learning curve.

The data sets with 303 records are further split with 165 and 138 for testing purposes whose discrimination threshold plots with 100 trials show the precision-recall and f1 score plots with training and testing unseen data sets show the best fits at 84% whose cross-validation might within +-10 scored. Similarly, the accuracy scored when 12 iterations mean squared scored 80.2 %, a similar result. After using random forest error

ISSN: 2319-7900 www.ijact.org Volume 13, Issue 1, February 2024

and cross-validation curves, 85% with 16 features scored optimal when five features were folded in each step. The data sets with 303 records are further split with 165 and 138 for testing purposes whose discrimination threshold plots with 100 trials show the precision-recall and f1 score plots with training and testing unseen data sets show the best fits at 84% whose cross-validation might within +-10 scored. Similarly, the accuracy scored when 12 iterations mean squared scored 80.2 %, a similar result. After using random forest error and cross-validation curves, 85% with 16 features scored optimal when five features were folded in each step. The Support Vector Machine attained an accuracy rate of 86%. Additionally, its Area Under the Curve (AUC) for predicting the absence of heart disease reached 92%. Notably, the macro average accuracy saw an increase to 93%. Similarly, the K-Nearest Neighbors model demonstrated improved classification, with a macro average accuracy of 93%. The prediction accuracy reached 89% for identifying cases with heart disease and 81% for instances without heart disease.

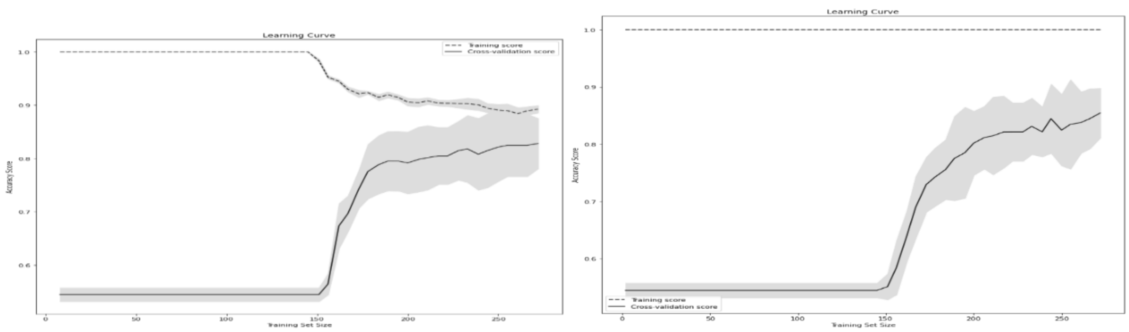

Fig 7: Learning curve k-nearest (a) and Random Forest(b)

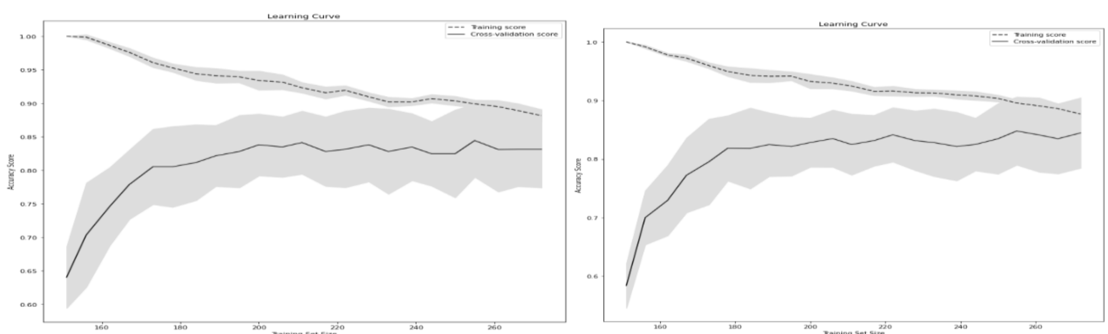

Fig 8: Learning curve of support vector(a) and Logistic regression(b)

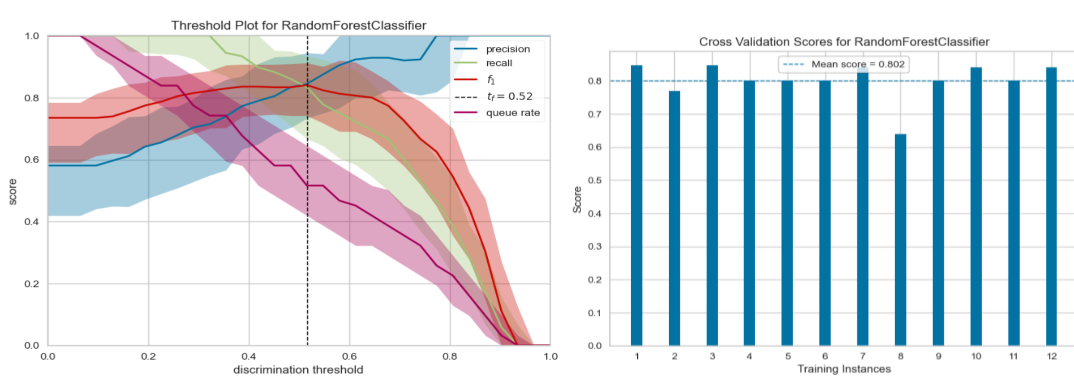

Fig 9: Accuracy score and threshold (a) Accuracy at 12 iterations (b)

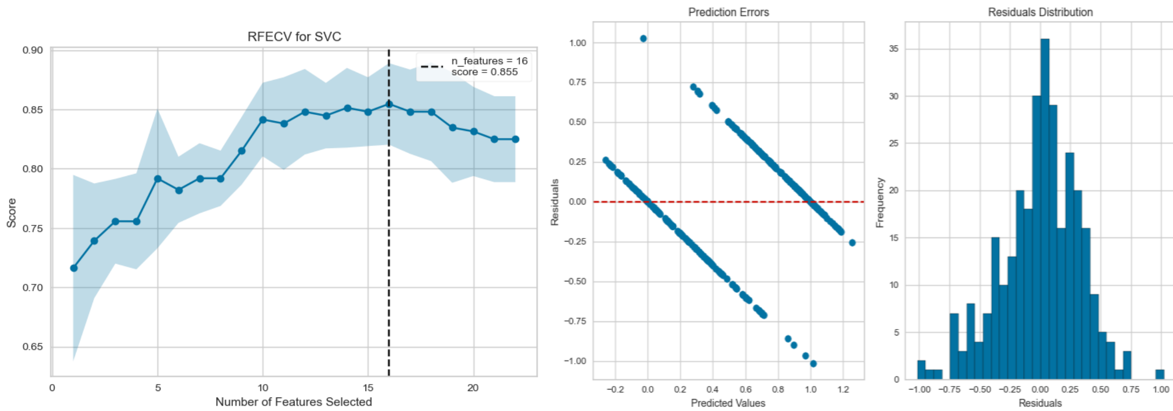

Fig10: Prediction error (a) and residual plot(b) histogram(c)

x

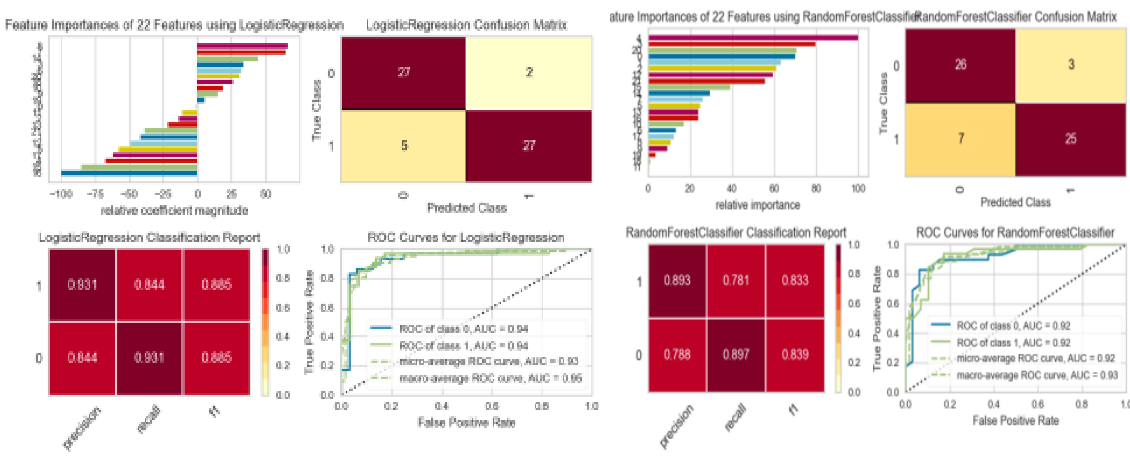

Fig11: Logistic regression (a) Random Forest summary statists (b)

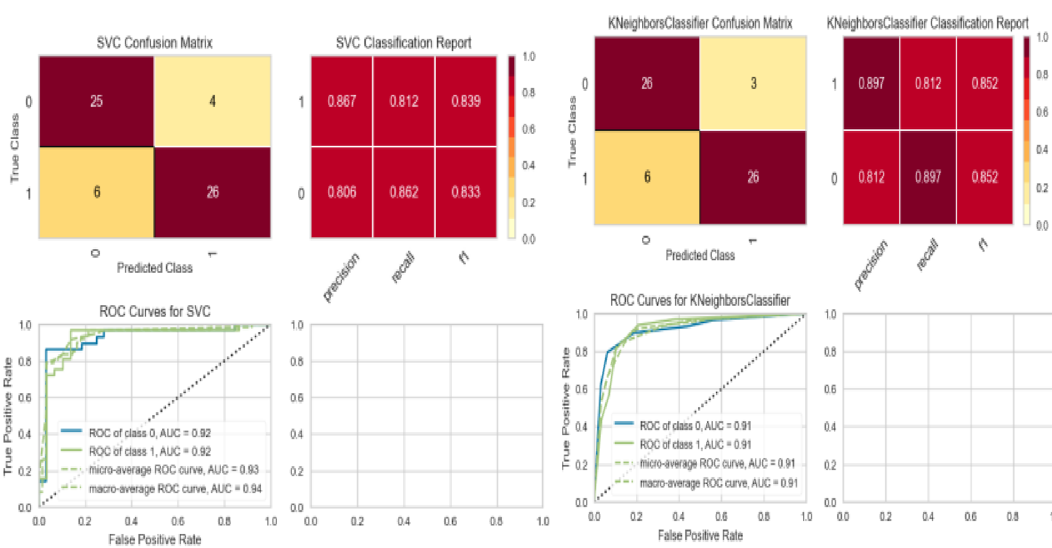

Fig12: Support vector (a) and K nearest summary statists(b)

Conclusion

Cross-validation is a statistical technique of evaluating algorithms by dividing facts into two segments, one used to analyze or train a version and the other used to train a version. Therefore, it's concluded that the k nearest model prediction accuracy ranges 5 to 13% higher than the default model with move validation model accuracy. Furthermore, while more than one fashion went for walks, the random forest model produced 90 % more accuracy than linear regression (81%) assist vector device and k nearest system gaining knowledge of version accuracy. The training samples used in machine learning models significantly influence the accuracy of heart disease prediction. Incorporating machine learning enables enhancement in accuracy compared to standalone model validations. Nevertheless, validating the hyperparameters of the training and test samples is essential to ensure the device achieves optimal accuracy for further research.

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this paper.

Data Availability: The open-source heart disease dataset containing 13 features is freely accessible at the following link: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset. The Python source code for migrating the source data to research data is also available in my GitHub repository: https://github.com/yagyarmal/Automl.git.

References

[1]. Amarbayasgalan, Tsatsral, et al. "An Efficient Prediction Method for Coronary Heart Disease Risk Based on Two Deep Neural Networks Trained on Well-Ordered Training Datasets." *IEEE Access*, vol. 9, 2021, pp. 135210–23.

[2]. Ansari, Mohd Faisal, et al. "A Prediction of Heart Disease Using Machine Learning Algorithms." *Image Processing and Capsule Networks*, edited by Joy Iong-Zong Chen et al., vol. 1200, Springer International Publishing, 2021, pp. 497–504. *DOI.org (Crossref)*, https://doi.org/10.1007/978-3-030-51859-2_45.

[3]. Anuradha, C., and T. Velmurugan. "A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance." *Indian Journal of Science and Technology*, vol. 8, no. 15, July 2015. *DOI.org (Crossref)*, https://doi.org/10.17485/ijst/2015/v8i15/74555.

[4]. Arora, Sarah, et al. "Diet and Lifestyle Impact the Development and Progression of Alzheimer's Dementia." *Frontiers in Nutrition*, vol. 10, 2023. *Google Scholar*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10344607/.

[5]. Barhoom, Ali MA, et al. *Prediction of Heart Disease Using a Collection of Machine and Deep Learning Algorithms*. 2022. *Google Scholar*, https://philpapers.org/rec/BARPOH-4.

[6]. Belkin, Mikhail, et al. "Reconciling Modern Machine-Learning Practice and the Classical Bias–Variance Trade-Off." *Proceedings of the National Academy of Sciences*, vol. 116, no. 32, 32, 2019, pp. 15849–54.

[7]. Benjamin, Emelia J., et al. "Heart Disease and Stroke Statistics—2019 Update: A Report from the American Heart Association." *Circulation*, vol. 139, no. 10, 2019, pp. e56–528.

[8]. Chowdhury, Rajiv, et al. "Dynamic Interventions to Control COVID-19 Pandemic: A Multivariate Prediction Modelling Study Comparing 16 Worldwide Countries." *European Journal of Epidemiology*, vol. 35, no. 5, 2020, pp. 389–99.

[9]. Dharma, Faisal, et al. "Prediction of Indonesian Inflation Rate Using Regression Model Based on Genetic Algorithms." *Jurnal Online Informatika*, vol. 5, no. 1, 2020, pp. 45–52.

[10]. Dodge, Jesse, et al. *Expected Validation Performance and Estimation of a Random Variable's Maximum*. arXiv:2110.00613, arXiv, 1 Oct. 2021. *arXiv.org*, http://arxiv.org/abs/2110.00613.

[11]. Gimenez-Nadal, Jose Ignacio, et al. "Resampling and Bootstrap Algorithms to Assess the Relevance of Variables: Applications to Cross Section Entrepreneurship Data." *Empirical Economics*, vol. 56, no. 1, 2019, pp. 233–67.

[12]. Hussain, Adedoyin A., and Kamil Dimililer. "Student Grade Prediction Using Machine Learning in IoT Era." *International Conference on Forthcoming Networks and Sustainability in the IoT Era*, Springer, 2021, pp. 65–81.

[13]. Kaur, Gaganjot, and Amit Chhabra. "Improved J48 Classification Algorithm for the Prediction of Diabetes." *International Journal of Computer Applications*, vol. 98, no. 22, July 2014, pp. 13–17. *DOI.org (Crossref)*, https://doi.org/10.5120/17314-7433.

[14]. Kernbach, Julius M., and Victor E. Staartjes. "Foundations of Machine Learning-Based Clinical Prediction Modeling: Part II—Generalization and Overfitting." *Machine Learning in Clinical Neuroscience*, edited by Victor E. Staartjes et al., vol. 134, Springer International Publishing, 2022, pp. 15–21. *DOI.org (Crossref)*, https://doi.org/10.1007/978-3-030-85292-4_3.

[15]. Khan, Anupam, and Soumya K. Ghosh. "Student Performance Analysis and Prediction in Classroom Learning: A Review of Educational Data Mining Studies." *Education and Information Technologies*, vol. 26, no. 1, Jan. 2021, pp. 205–40. *DOI.org (Crossref)*, https://doi.org/10.1007/s10639-020-10230-3.

[16]. Mahesh, T. R., et al. "The Stratified K-Folds Cross-Validation and Class-Balancing Methods with High-Performance Ensemble Classifiers for Breast Cancer Classification." *Healthcare Analytics*, vol. 4, 2023, p. 100247.

[17]. Maldonado, Sebastián, et al. "Out-of-Time Cross-Validation Strategies for Classification in the Presence of Dataset Shift." *Applied Intelligence*, vol. 52, no. 5,

Mar. 2022, pp. 5770–83. *DOI.org (Crossref)*, https://doi.org/10.1007/s10489-021-02735-2.

[18]. Mohan, Senthilkumar, et al. "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques." *IEEE Access*, vol. 7, 2019, pp. 81542–54.

[19]. Nadar, Nityashree, and R. Kamatchi. "A Novel Student Risk Identification Model Using Machine Learning Approach." *Int. J. Adv. Comput. Sci. Appl*, vol. 9, 2018, pp. 305–09.

[20]. Naz, Huma, and Sachin Ahuja. "Deep Learning Approach for Diabetes Prediction Using PIMA Indian Dataset." *Journal of Diabetes & Metabolic Disorders*, vol. 19, no. 1, 2020, pp. 391–403.

[21]. Olaniyi, Ebenezer O., et al. "In-Line Grading System for Mango Fruits Using GLCM Feature Extraction and Soft-Computing Techniques." *International Journal of Applied Pattern Recognition*, vol. 6, no. 1, 2019, pp. 58–75.

[22]. Schmidt, Jonathan, et al. "Recent Advances and Applications of Machine Learning in Solid-State Materials Science." *Npj Computational Materials*, vol. 5, no. 1, 2019, pp. 1–36.

[23]. Shukla, Nagesh, et al. "Breast Cancer Data Analysis for Survivability Studies and Prediction." *Computer Methods and Programs in Biomedicine*, vol. 155, 2018, pp. 199–208.

[24]. Smirani, Lassaad K., et al. "Using Ensemble Learning Algorithms to Predict Student Failure and Enabling Customized Educational Paths." *Scientific Programming*, edited by Chenxi Huang, vol. 2022, Apr. 2022, pp. 1–15. *DOI.org (Crossref)*, https://doi.org/10.1155/2022/3805235.

[25]. Touzani, Samir, et al. "Gradient Boosting Machine for Modeling the Energy Consumption of Commercial Buildings." *Energy and Buildings*, vol. 158, 2018, pp. 1533–43.

[26]. Townsend, Nick, et al. "Epidemiology of Cardiovascular Disease in Europe." *Nature Reviews Cardiology*, vol. 19, no. 2, 2, 2022, pp. 133–43.

[27]. Usama, Mohd, et al. "Self-Attention Based Recurrent Convolutional Neural Network for Disease Prediction Using Healthcare Data." *Computer Methods and Programs in Biomedicine*, vol. 190, 2020, p. 105191.

[28]. Xiong, Biao, et al. "Semi-Supervised Classification Considering Space and Spectrum Constraint for Remote Sensing Imagery." *2010 18th International Conference on Geoinformatics*, IEEE, 2010, pp. 1–6.

[29]. Ye, Zheng, et al. "Predicting Beneficial Effects of Atomoxetine and Citalopram on Response Inhibition in Parkinson's Disease with Clinical and Neuroimaging Measures." *Human Brain Mapping*, vol. 37, no. 3, 2016, pp. 1026–37.

[30]. Yousafzai, Bashir Khan, et al. "Application of Machine Learning and Data Mining in Predicting the Performance of Intermediate and

Secondary Education Level Student." *Education and Information Technologies*, vol. 25, no. 6, 2020, pp. 4677–97.

[31]. Zuhair, Mohamed, et al. "Estimation of the Worldwide Seroprevalence of Cytomegalovirus: A Systematic Review and Meta-Analysis." *Reviews in Medical Virology*, vol. 29, no. 3, 2019, p. e2034.