Project Title:
Comparative analysis on Cardiovascular Disease (CVD) Prediction with Feature Subset Selection

Presented by **Ishrat Tanzila Farah**
Student ID: 1017052063
Group ID: 4

Presentation Outline

- Motivation
- Dataset
- Related Works
- Proposed Methods
- Progress
- Result Analysis
- Future Work

Motivation

- CVD is a common death factor worldwide
- Many features of heart datasets, which contain relevant as well as irrelevant and redundant features
- Redundant features deteriorates performance
- Removing those features before applying classifier techniques is necessary
- Many existing hybrid models
- **Scopes for improved performance exists**
- **Many feature subset selection methods not explored yet**

Dataset

Source: UCI Machine Learning Repository
Number of Features: 13

Target Attribute: Represents the diagnosis of heart disease with 5 values.
0 = absence
1 to 4 = presence → 1

Missing Values: 6 [Thal : 2, CA: 4]

↓

Imputed by Most Frequently Occuring Value

1. Sex (value 1: Male; value 0 : Female)
2. Chest Pain Type (value 1: typical type 1 angina, value 2: typical type angina, value 3: non-angina pain; value 4: asymptomatic)
3. Fasting Blood Sugar (value 1: > 120 mg/dl; value 0: < 120 mg/dl)
4. Restecg – resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
5. Exang – exercise induced angina (value 1: yes; value 0: no)
6. Slope – the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: downsloping)
7. CA – number of major vessels colored by floursopy (value 0 – 3)
8. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
9. Trest Blood Pressure (mm Hg on admission to the hospital)
10. Serum Cholesterol (mg/dl)
11. Thalach – maximum heart rate achieved
12. Oldpeak – ST depression induced by exercise relative to rest
13. Age in Year

Already Explored Feature Subset Selection Methods

- Logistic Regression (LR)
- Multivariate Adaptive Regression Splines (MARS)
- Rough Set (RS) techniques with Firefly Algorithm
- PCA
- Chi Square
- Wavelets
- Genetic Algorithm
- Mean Fisher score-based feature selection algorithm
    - Forward Feature Selection Algorithm
    - Reverse Feature Selection Algorithm
- Sequential Forward Selection
- Sequential Backward selection
- Symmetrical Uncertainty (rank the attributes Information Gain)
- Brute Force with minimum 3 attributes to find feature subset

Related Works

Hybrid intelligent modeling schemes for heart disease classification by Yuehjen E. Shao , Chia-Ding Hou, Chih-Chou Chiu

- **Proposed Method: Hybrid Model**
  - Logistic Regression (LR)
  - Multivariate Adaptive Regression Splines (MARS)
  - Artificial Neural Network (ANN)
  - Rough Set (RS) techniques

- **Feature Subset Selection:**
  - LR [Wald forward method to recognize explanatory variables)]
  - MARS [Least contributions were deleted using generalized cross validation (GCV)]
  - RS techniques
- **Classification:** ANN, LR, MARS, RS

- **Outcome:** Outperform the typical, single-stage ANN method

Highest Accuracy: MARS-LR (83.93%)
Reduced: 7 Features

AIR comparisons for single-stage and hybrid models.

Models	Explanatory variables	AIR (%)
ANN alone	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}$	76.79
RS alone	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}$	78.60
MARS alone	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}$	78.57
LR–ANN	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_9, X_{10}, X_{11}, X_{12}, X_{13}$	78.57
MARS–ANN	$X_2, X_3, X_4, X_{10}, X_{12}, X_{13}$	82.14
RS–ANN	$X_2, X_3, X_4, X_6, X_7, X_9, X_{10}, X_{11}, X_{12}, X_{13}$	79.50
LR–MARS	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_9, X_{10}, X_{11}, X_{12}, X_{13}$	78.57
RS–MARS	$X_2, X_3, X_4, X_6, X_7, X_9, X_{10}, X_{11}, X_{12}, X_{13}$	80.36
LR–RS	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_9, X_{10}, X_{11}, X_{12}, X_{13}$	81.25
MARS–RS	$X_2, X_3, X_4, X_{10}, X_{12}, X_{13}$	76.79
MARS–LR	$X_2, X_3, X_4, X_{10}, X_{12}, X_{13}$	83.93
RS–LR	$X_2, X_3, X_4, X_6, X_7, X_9, X_{10}, X_{11}, X_{12}, X_{13}$	83.93

Classification of Healthcare Data using Genetic Fuzzy Logic System and Wavelets by Nguyen, Abbas, Douglas, and Saeid

- **Proposed Method:** GSAM+Wavelets
- **Feature Subset Selection:** Wavelets
- **Classification:** An integration of fuzzy standard additive model (SAM) with genetic algorithm (GA), called GSAM
- **Motivation behind using Wavelets**
  - In PCA, eigenvectors corresponding to the largest variance of the data are selected
  - But these directions do not necessarily provide the best separation of the classes
  - There is a possibility that the information for separating the clusters is represented in principal components with low eigenvalues, which are often ignored.
  - The use of WT is thus promoted in this research.
- **Outcome:** Achieved higher accuracy than feature reduction with PCA

Wavelet Transformation (WT)

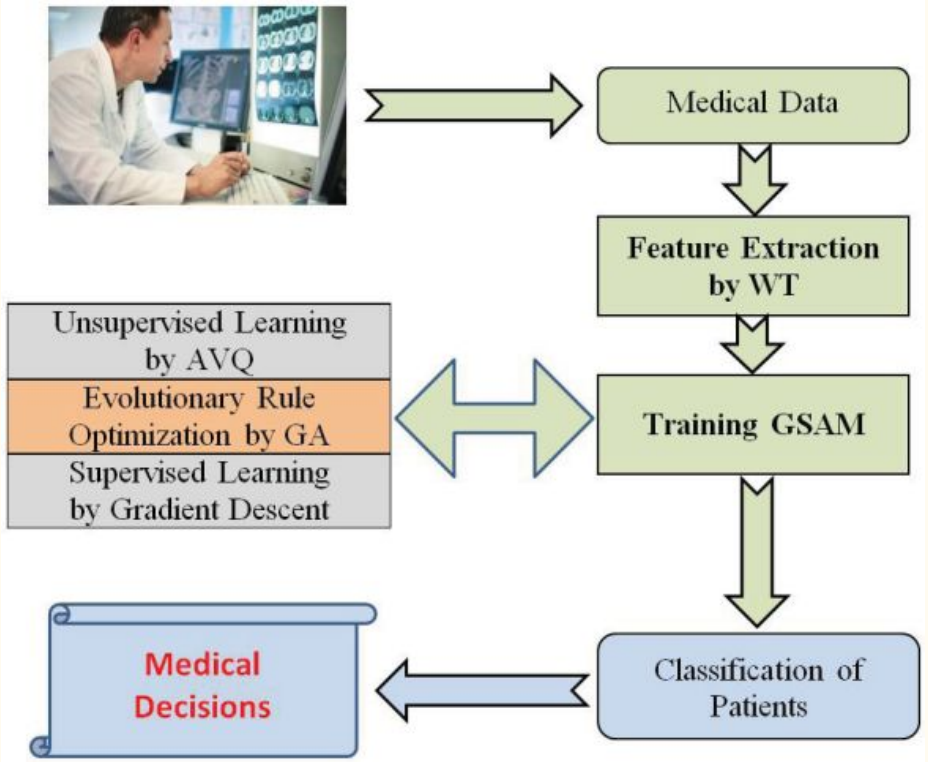
- WT represents a signal in a time-frequency fashion.
- Once the mother wavelet **φ**(x) is fixed, translations and dilations of the mother wavelet can be formed with the equation:

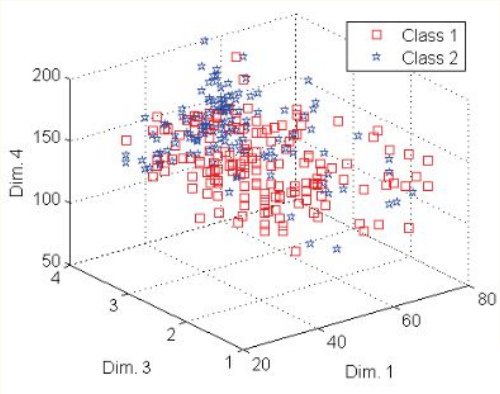
$$\left\{\varphi\left(\frac{x-b}{a}\right), (a,b) \in \mathrm{R}^{+} \times \mathrm{R}\right\}$$

- Employed four-level decomposition using Haar wavelets, due to their compact support and orthogonality
- Selected coefficients should have the largest deviation from normality. For this reason, the Lilliefors modification of a Kolmogorov-Smirnov (KS) test for normality is employed.
- The test compares the cumulative distribution function of the data F(x) with a Gaussian distribution function G(x)given a dataset x.
- Deviation from normality is measured by max|F(x)-G(x)|.

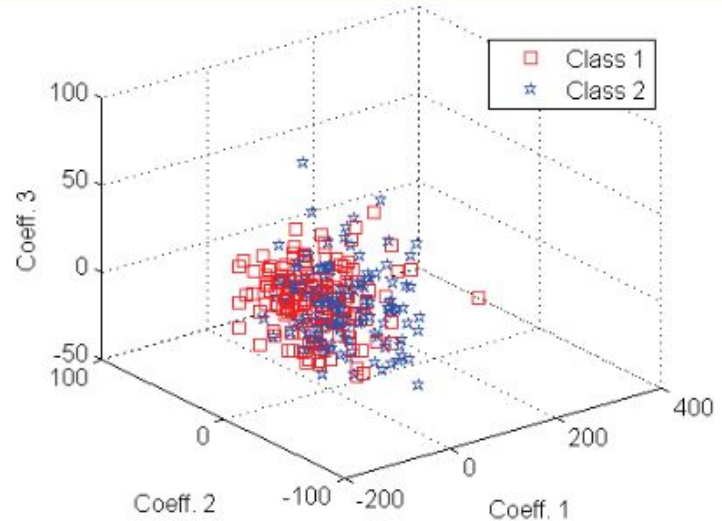
Continued

- Wavelet transformation is employed to extract discriminative features
- GSAM learning process comprises three continual steps:
  - Rule initialization by unsupervised learning using the adaptive vector quantization clustering,
  - Evolutionary rule optimization by GA
  - Parameter tuning by the gradient descent supervised learning.

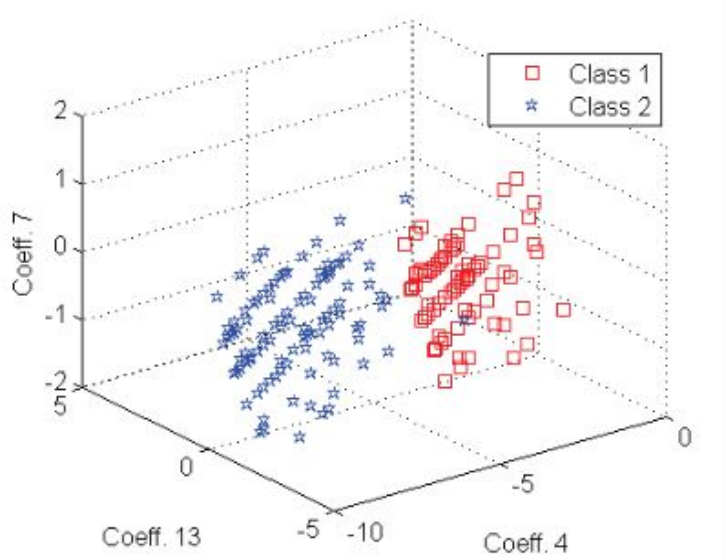




Original features



Projection of 3 features extracted by PCA



Projection of 3 wavelet features

Highest Accuracy: GSAM+Wavelets (78.78%)

Methods	Accuracy%		
	Original	PCA	Wavelets
PNN	55.06	55.10	73.80
SVM	57.25	61.10	74.27
FARTMAP	62.58	59.87	63.46
ANFIS	73.10	64.25	74.90
SAM	60.81	61.36	75.35
GSAM	63.17	63.59	78.78

A highly accurate firefly based algorithm for heart disease prediction by Nguyen , Phayung, Herwig

- **Proposed Method:** Rough sets based attribute reduction and interval type-2 fuzzy logic system (IT2FLS)
- **Feature Subset Selection:** The rough sets based attribute reduction using Firefly Algorithm
- **Classification:** IT2FLS - a hybrid learning process comprising fuzzy c-mean clustering algorithm and parameters tuning by chaos firefly and genetic hybrid algorithms.
- **Limitations**
  - The proposed rough sets based attribute reduction is unmanageable when the number of attributes is huge or when the number of records is large.
  - The training time of interval type-2 fuzzy logic system by chaos firefly and genetic hybrid algorithms is quite slow.

The Firefly Algorithm (FA)

- Swarm Intelligence Technique: Stochastic, Meta-heuristic algorithm
- **This algorithm is inspired by the flashing light of fireflies in nature.**
- The main ideas of the firefly algorithm is interpreting light intensity characteristics as follows:
  - All fireflies are unisex and there may be an attractive in any two fireflies.
  - Their attractiveness is proportional to their light intensity.
  - A firefly with lower light intensity will move toward the fireflies with higher light intensity.
  - If there is not firefly with higher light intensity, the firefly will randomly move in search space.
  - The light intensity of a firefly is determined by fitness function.

Attraction function: $\beta = \beta_0 \times e^{-\gamma r2ij}$

r = Distance between two fireflies i and j respectively
β_0 = An attraction parameter that is the attractiveness at r = 0
γ = Light absorption coefficient

Movement of fireflies:

$$X_i(t+1) = X_i(t) + \beta(X_i(t) - X_j(t)) + \alpha\left(rand - \frac{1}{2}\right)$$

X i (t) ,X j (t) = Positions of firefly i with lower light intensity and firefly j with higher intensity at time t respectively
α = Random parameter which determines random behavior of movement of fireflies

Reduced 9 Attributes, Accuracy: 88.3%

BPSORS-AR: Binary Particle Swarm Optimization & Rough Sets

CFARS-AR: Chaos Firefly Algorithm & Rough Sets (Proposed Method)

Models	Without attribute reduction	BPSORS-AR	CFARS-AR
Naive Bayes	83.3	79.6	85.2
SVM	75.9	75.9	81.5
ANN	77.8	74.1	81.5
New Approach	86	87.0	88.3

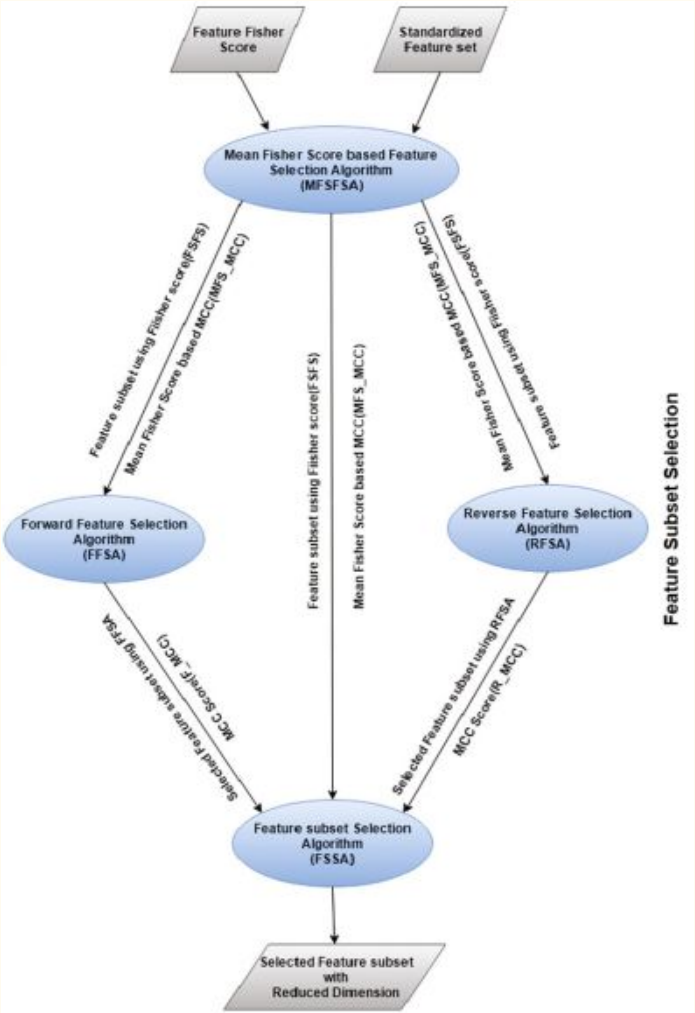
Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using SVM by S.M.Saqlain, M.Sher, F. A. Shah, Imran Khan, M.U. Ashraf, M. Awais, Anwar Ghani

Feature Subset Selection:
(1) mean Fisher score-based feature selection algorithm
(2) forward feature selection algorithm
(3) reverse feature selection algorithm

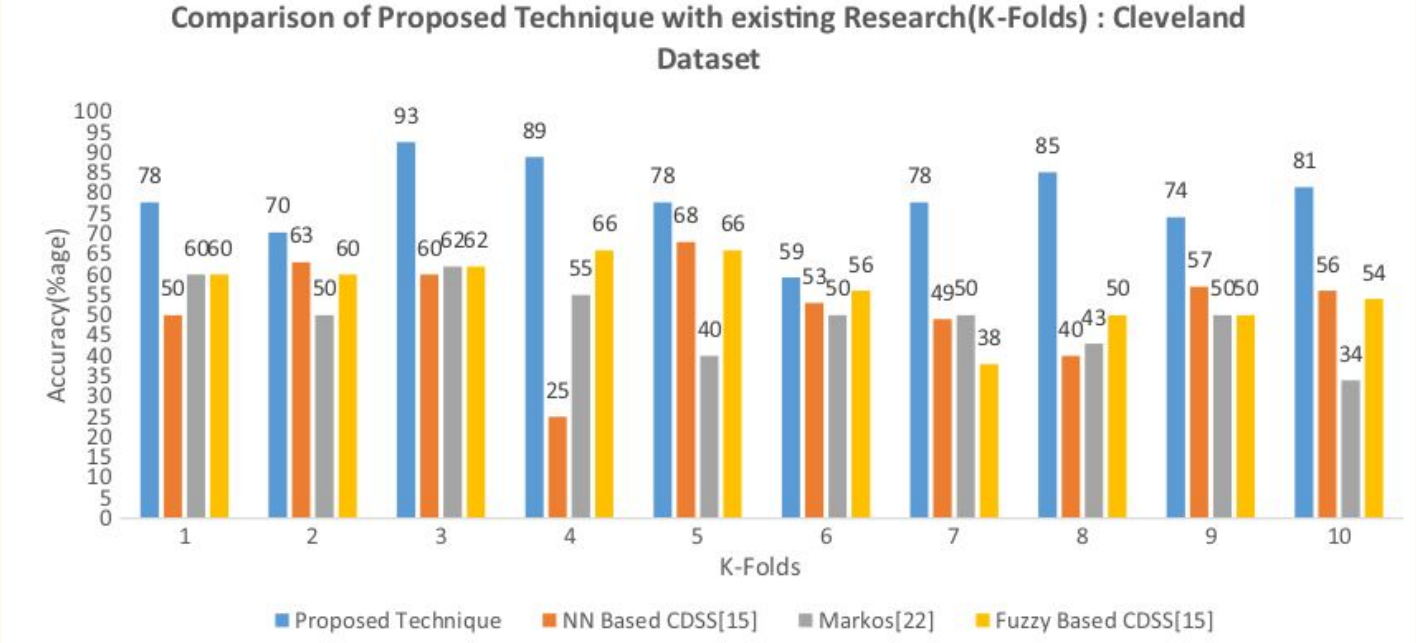
Classification: RBF kernel-based SVM

Outcome: Achieved considerably better results than some of the other existing techniques.

- The proposed algorithms use individual Fisher scores of the features, along with MCC scores of the subsets, for their selection in the feature subsets.
- The proposed feature selection algorithms select one feature subset each and among all of them the feature subset with the **higher MCC score and lower dimension** is a good choice of selection.



Feature subset dimension through MFSFSA	Feature subset dimension through FFSA	Feature subset dimension through RFSA	Feature subset through FSSA	Classification accuracy (%)
6	7	7	FFSA	81.19



Identification of significant features and data mining techniques in predicting heart disease. Authors: Mohammad Shafenoor Amin 1,3 , Yin Kia Chiam 1 , Kasturi Dewi Varathan 2

Proposed Method: Prediction models using different combination of features, and classification techniques.

Feature Subset Selection: Brute Force method having lower bound with 3 features
(total of 8100 combinations of the features)

7 Classification Techniques: k-NN, Decision Tree, Naive Bayes, Logistic Regression, Support Vector Machine, Neural Network and Vote (i.e. a hybrid technique with Naïve Bayes and Logistic Regression)

Outcome: VOTE + 9 Significant Features (Accuracy: 87.41%)

Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection by M.A. Jabbar, B.L Deekshatulu & P. Chandra

Proposed Method: PCA+ANN, χ^2+ANN, GA+ANN

Feature Subset Selection: PCA , Chi Square, GA

Classification: ANN

Dataset: Andhra Pradesh heart disease dataset

Outcome: 100% Accuracy

Data set	PCA	Chi square	GA
Weather	100	100	100
Pima	98.82	98.82	78.3
hypothyroid	97.06	97.64	97.37
breast cancer	97.9	97.64	95.45
liver disorder	95.07	70	84.63
primary tumor	80	83.18	83.18
heart stalog	98.14	97.7	99.62
lymph	99.32	100	99.32
heart disease A.P	100	100	100
average	**96.2**	**93.8**	**93.09**

Heart Disease Classification using Nearest Neighbor Classifier with Feature Subset Selection. Authors: M.A. Jabbar, B.L. Deekshatulu, P. Chandra

Proposed Method: KNN+SU

Feature Subset Selection: Symmetrical Uncertainty of attribute (a measure which compensates Information Gain)

Classification: K Nearest Neighbor

Dataset: Andhra Pradesh heart disease dataset

Outcome: 100% Accuracy

Achieved same accuracy as ANN+PCA, ANN+χ^2, ANN+GA for Heart Disease Dataset (Accuracy: 100%)

Table 7. Classification accuracy of (KNN+SU) with (ANN+PCA) and (ANN+χ^2)

Data set name	Our approach	ANN+PCA	ANN+χ^2
Liver disorder	100	95.07	70
Diabetes	100	98.82	98.8
lympography	100	99.32	100
Primary tumor	61.35	80	83.18
Heart stalog	100	98.14	97.7
Breast cancer	96.85	97.9	97.64
Heart disease	100	100	100
Average	94.02	95.60	92.39

Feature Subset Selection Methods Not Explored Yet

Feature Subset Selection Methods to be applied

- Linear Discriminant Analysis (LDA)
- Recursive Feature Elimination (RFE)
- Gaussian Random Projection (GRP)
- Sparse Random Projection (SRP)
- Mutual information (MI)
- F-test
- L1-Norm based feature selection

Classifiers to be used

- Logistic Regression
- Support Vector Machine
- Decision Tree Classifier
- Random Forest Classifier
- Naive Bayesian Classifier

Progress till now

- Logistic Regression ✔️
- Support Vector Machine
- Decision Tree Classifier
- Random Forest Classifier
- Naive Bayesian Classifier

Validation: Repeated Stratified K-Fold Cross Validation

Performance Measures: Accuracy, Precision, Recall, F1 score, ROC_AUC

Logistic Regression without Feature Subset Selection

Accuracy	83.95%
Precision	87.60%
Recall	76.30%
F1-Score	81.1%
ROC_AUC	90.5%

Correlation Based Feature Subset

Spearman

Kendall

Pearson

Correlation Based Feature Subset

	pearson	kendall	spearman
cp	0.414446	0.440972	0.472006
thalach	0.417167	-	0.423467
exang	0.431894	0.431894	0.431894
oldpeak	0.424510	-	0.413382

Correlation Method	#Features
pearson	4
kendall	2
spearman	4

Threshold: 0.4

Accuracy: 79.09%

Recursive Feature Elimination (RFE)

- Given an external estimator that assigns weights to features, recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features.
- First, the estimator is trained on the initial set of features and the importance of each feature is obtained through their corresponding coefficients.
- Then, the least important features are pruned from the current set of features.
- This procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.

Feature Subset with RFE

#Features	Feature_Subset	accuracy	precision	recall	f1	roc_auc
7	['sex', 'cp', 'thalach', 'exang', 'oldpeak', 'ca', 'thal']	85.16%	89.04%	**77.69%**	**82.60%**	90.49%
11	['sex', 'cp', 'trestbps', 'fbs', 'restecg', 'thalach', 'exang', 'oldpeak', 'slope', 'ca', 'thal']	**85.27%**	**90.20%**	76.78%	82.50%	**90.64%**

Linear Discriminant Analysis (LDA)

- A classifier with a linear decision boundary, generated by fitting class conditional densities to the data using Bayes' rule.
- LDA seeks to best separate (or discriminate) the samples in the training dataset by their class value.
- Specifically, the model seeks to find a linear combination of input variables that achieves the maximum separation for samples between classes (class centroids or means) and the minimum separation of samples within each class.
- LDA can be used to calculate a projection of a dataset and select a number of dimensions or components of the projection to use as input to a model.
- The model fits a Gaussian density to each class, assuming that all classes share the same covariance matrix.

Feature Subset with LDA

Number of components (<= min(n_classes - 1, n_features))

n_classes = 2

n_features = 13

Number of components <= min(2 - 1, 13)

<= min(1, 13) = **1**

accuracy	precision	recall	f1	roc_auc
86.05%	90.25%	78.44%	83.55%	92.08%

L1-norm based Feature Subset Selection

- Linear SVC penalized with the L1 norm
- Have sparse solutions: many of their estimated coefficients are zero.
- When the goal is to reduce the dimensionality of the data to use with another classifier, they can be used to select the non-zero coefficients.

#Features	accuracy	precision	recall	f1	roc_auc
6	84.40%	87.53%	77.69%	81.91%	89.68%

Feature Subset with Gaussian Random Projection

- Gaussian Random Projection reduces the dimensionality by projecting the original input space on a randomly generated matrix where components are drawn from the following distribution:

$$N(0, \frac{1}{n_{components}})$$

#Components	accuracy	precision	recall	f1	roc_auc
9	81.08%	80.03%	78.86%	79.10%	88.22%

Sparse Random Projection

Sparse Random Projection reduces the dimensionality by projecting the original input space using a sparse random matrix.

Sparse random matrices are an alternative to dense Gaussian random projection matrix that guarantees similar embedding quality while being **much more memory efficient and allowing faster computation** of the projected data.

If we define s = 1 / density, the elements of the random matrix are drawn from

$$\begin{cases} -\sqrt{\frac{s}{n_{\text{components}}}} & & 1/2s \\ 0 & \text{with probability} & 1 - 1/s \\ +\sqrt{\frac{s}{n_{\text{components}}}} & & 1/2s \end{cases}$$

Where $n_{\text{components}}$ is the size of the projected subspace.

Feature Subset with Sparse Random Projection

#Components	accuracy	precision	recall	f1	roc_auc
5	80.31%	80.93%	75.57%	77.71%	87.74%
8	80.56%	80.38%	77.25%	78.36%	87.33%
9	80.65%	83.7%	72.18%	77.15%	86.96%

Feature Subset using Mutual Information (MI)

- Mutual information (MI) between two random variables is a non-negative value, which measures the dependency between the variables.
- It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency.
- The function relies on nonparametric methods based on entropy estimation from k-nearest neighbors distances.
- It can be used for univariate features selection.

Feature Subset using Mutual Information (MI)

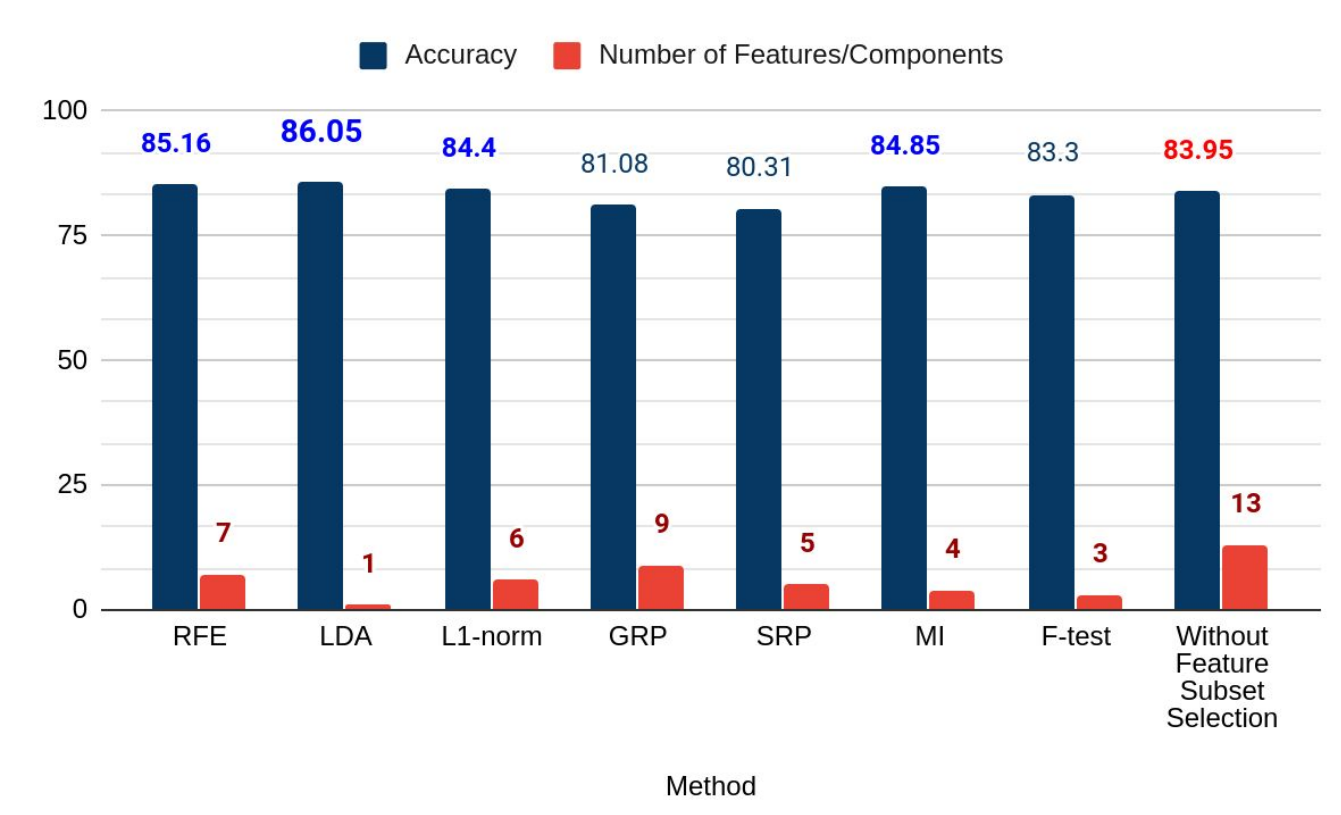
#Features	accuracy	precision	recall	f1	roc_auc
4	**84.85%**	**87.59%**	**79.18%**	**82.70%**	89.41%
5	82.09%	84.43%	78.19%	79.96%	89.48%
6	83.30%	84.52%	78.64%	80.75%	89.17%
7	83.52%	85.96%	77.93%	80.68%	89.13%
8	82.42%	85.14%	78.19%	81.09%	88.97%
9	**83.85%**	85.38%	78.41%	81.50%	**89.78%**

Feature Subset using F-Test

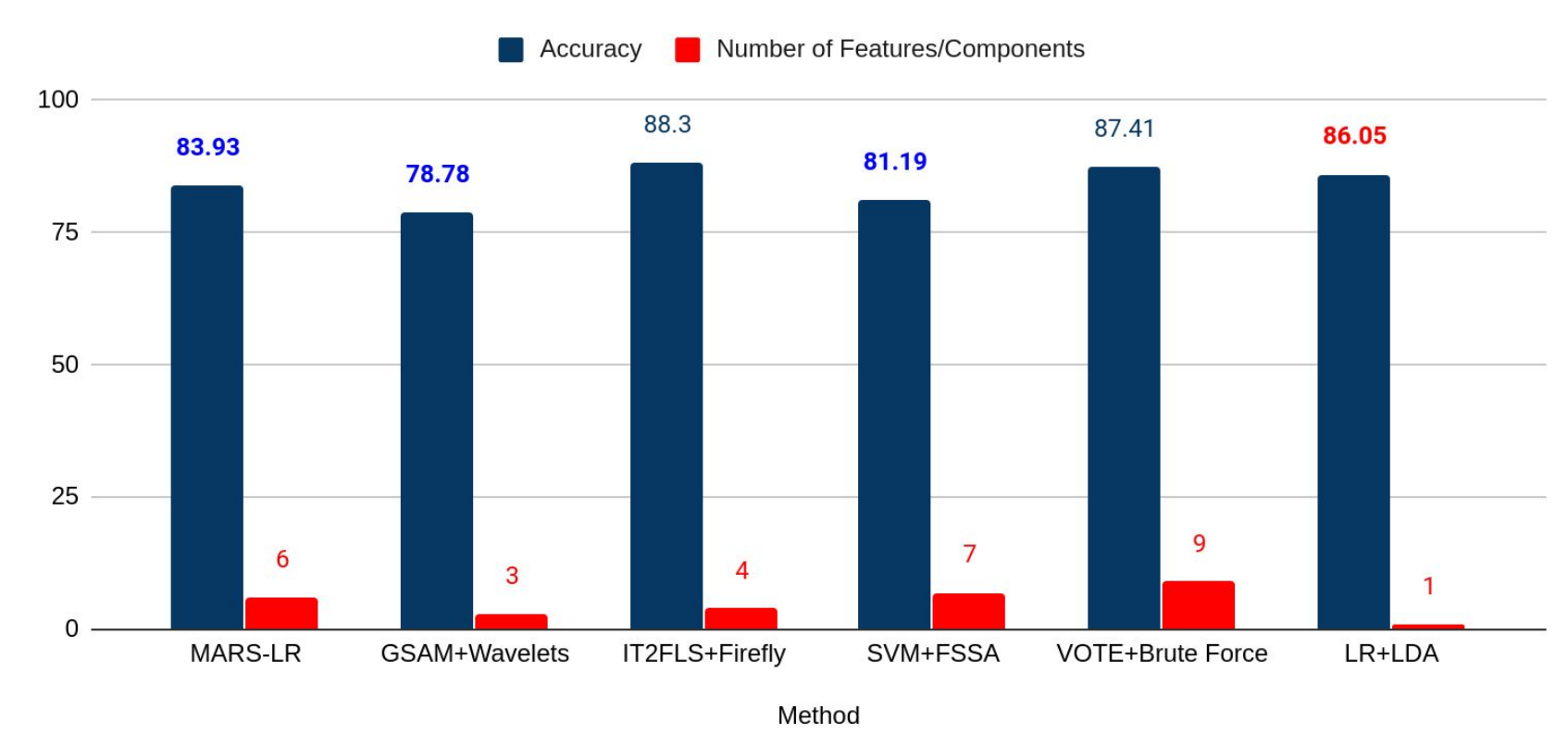
Computes the ANOVA F-value for the provided sample.

#Features	accuracy	precision	recall	f1	roc_auc
3	**83.30%**	88.00%	74.80%	80.31%	88.35%
4	82.29%	82.88%	76.43%	79.11%	88.88%
5	**83.51%**	85.97%	77.45%	81.03%	89.43%
6	**84.40%**	87.53%	77.69%	81.91%	89.68%
7	**84.49%**	85.32%	78.21%	81.23%	89.44%
8	**84.94%**	85.20%	79.16%	81.67%	89.99%
9	84.28%	85.99%	79.63%	82.24%	90.13%
10	84.61%	85.42%	78.92%	81.72%	89.62%

Accuracies of different feature subset selection methods applied



Comparative accuracy with existing hybrid methods



Future Work

- Apply other classifiers
- Runtime comparisons

Thank You

Any Question/Suggestion?