Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 46 No. 2 (2025)

Revolutionizing Heart Disease Diagnosis with Superior Feature Selection: The Power of mRMR

[1]Rajani Alugonda, [2]Satya Prasad Kodati

[1]Research Scholar, Assistant Professor, Department of Electronics and Communication Engineering, JNTUK, Kakinada, Andhra Pradesh, India.

[2]Professor, Department of Electronics and Communication Engineering, JNTUK, Kakinada, Andhra Pradesh, India.

Abstract: A serious concern to human and their health upcoming years is heart disease. To prompt diagnosis care, patients frequently experience impairment or even pass away. The diagnosis is directly dased on the experience of the many doctors, and the situation is made worse by the numerous issues related to heart disease that place a great burden on them. Therefore, it makes sense to introduce computer-aided approaches to help doctors diagnose cardiac disease in order to improve treatment. Nowadays, researchers typically use the feature selection approach to the processed (13 features) dataset that was chosen by physicians. This is improper because the feature size is so small. The usefulness of the unprocessed dataset is overlooked, and many are unaware that it may contain latent. The mRMR is better than previous approaches, and the incremental feature selection method works well. It has the least helpful features in addition to the best accuracy. On the Cleveland dataset, it has 100% accuracy with 8 features, on the Hungarian dataset, it has 98.3% accuracy with 14 features, and on the Long-beach-VA dataset, it has 99% accuracy with 9 characteristics. Additionally, we discover that certain characteristics—which physicians consider insignificant—have a role in classification and ought to catch their attention.

TERMS INDEX: Heart illness, mutual information, feature selection, and mRMR.

1. INTRODUCTION:

Due to their suddenness, cardiovascular diseases (CVDs) claim millions of lives each year. The patient is at risk of becoming disabled or maybe dying if treatment is delayed. Therefore, one successful strategy to save lives is to diagnose cardiac disease early and accurately .As medical research has advanced, physicians have identified numerous heart disease signs, directly created a more potent classifier by implementing the majority voting technique. Four machine learning techniques— K-Nearest Neighbour, Random Forest, and Logistic Regression—make up this ensemble approach. Out of the four algorithms, the one with the highest prediction accuracy is 88%. 90% accuracy is attained by the ensemble method utilizing the voting procedure. These approaches do not fully utilize the dataset and disregard the differences in the features. A fundamental step in the current study area is reducing the dimensionality of datasets, which entails removing redundant or unimportant information [11] that will drastically impair the effectiveness of machine learning techniques or raise computing costs. Feature extraction and feature selection are the two primary branches of dimensionality reduction. The two most used feature extraction techniques are PCA and LDA [12]. A neural network approach using PCA was presented by Karayilan et al. [13] to identify cardiac disease. Additionally, approach based on the LDA was proposed by Kolukisa et al. [14]. A technique for choosing PCA and as the input of Random Forest using spectroscopic data was presented by Shafizadeh-Moghadam and Hossein [15]. They discovered that the target variable's most pertinent primary components weren't the first.

A dataset's features that follow a specific algorithm are chosen through feature selection. These techniques preserve all feature information and improve the interpretability of prediction outcomes. Additionally,

it is an effective method for preventing overfitting [16], which is brought on by the expansion of feature dimensionality space. Feature selection, filter methods, wrapper methods, and embedding techniques are the three general categories [17]. Filter techniques assess their value or rank using a set of criteria. It's computationally.

. Elastic net plus a genetic algorithm make up this technique. The first layer uses the evolutionary algorithm, one of the heuristic techniques, to choose a local optimization feature subset. The second layer uses the penalty factors to remove the redundant features.

The well-known decision tree algorithm is one of the filter techniques based on the Shannon information entropy theory that have become significant in artificial intelligence since its inception [24]. The first purpose of the mutual information maximization (MIM) method is to reduce the uncertainty of class labels. Nevertheless, this approach only takes into account a feature's relevance and disregards its redundancy, resulting in the presence of redundancy in the chosen characteristics. The concept of feature redundancy is introduced by Peng et al. [25] to enhance the effect of mutual information, significantly increasing its application. A frequent occurrence in datasets related to heart disease is value missing, which significantly impairs feature selection techniques. In order to avoid the detrimental impact that missing values can have, the mRMR technique will disregard certain feature values.

1.1. OUR-CONTRIBUTIONS:

These shortcomings still exist in current efforts, which primarily use certain algorithms selection techniques of 13 features. The feature selection approach on so few features is not necessary. The dataset limits the algorithm's performance, and the 74-feature dataset ought to have more corroborating data. The mutual information approach is a useful option for taking into account the interaction within features. Additionally, the efficacy of the incremental feature combination approach needs to be confirmed. In this study, we eliminated the dataset's performance limitation and highlighted the significance of the 74 characteristics dataset. One of the filter methods, the mRMR approach.

The Random Forest method, mRMR, Kendall τ correlation, LDA, PCA, and LDA theories and algorithms are presented in Section II. Additionally, the flowchart for the entire experiment and datasets related to heart disease are introduced in this section. A thorough comparison of the findings, discussion, and conclusion are presented in Section III. and Section IV concludes by summarizing the findings of this effort and outlining a future plan.

1.2.METHODS AND DATASETS:

One method for reducing feature scale is dimensionality reduction [30]. It finds the latent information in addition to reducing their dimensionality and speeding up the training procedure. This type of approach can occasionally highlight the significance of certain features [31], [32]. The specifics of the Random Forest, mRMR, Kendall τ correlation, LDA, and PCA techniques will be covered in this section.

A) The PCA Method

PCA type of method that reflects the original data and investigates the most important factors. To accomplish the goal of dimensionality reduction, it employs matrix transformation, minimizes information loss, and allows variance values to grow as much as possible. Numerous research fields, including text mining and picture recognition, have made use of PCA.

2. Proposed Algorithms: The proposed algorithms are as follows:

Table 2 displays the distributions for Cleveland, Hungarian, and Long Beach, Virginia. Following the removal of any missing or worthless values, they have 280, 249, and 121 samples, respectively. Whereas the presence indicates illness, the absence indicates health. They're both out of balance.

Algorithm-1 Kendall
Output the new dataset D

```
def select_top_n_features(dataset, target_label, n):
    """
    Selects the top n features from the dataset based
 on Kendall's tau correlation with the target label.

    Parameters:
    dataset (pd.Data Frame): The input dataset containing features and the target label.
    target_label (str): The name of the target label column.
    n (int): The number of top features to select.

    Returns:
    pd.DataFrame:
 A new DataFrame containing the top n features and the target label.
    """
    # Separate features (X)
 and target label (y)
    X = dataset.drop(columns=[target_label])
    y = dataset[target_label]

    # Calculate Kendall's tau for each feature
    tau_values = {
        feature: abs(kendalltau(X[feature], y).correlation)
        for feature in X.columns
    }

    # Sort features by absolute tau value in descending order and select top n
    top_features = sorted(tau_values, key=tau_values.get, reverse=True)[:n]

    # Create a new DataFrame with the selected features and target label
    selected_data = dataset[top_features + [target_label]]
    return selected_data
```

ALGORITHM 2:

```
# Function to process dataset
```

```
def process_dataset(D, n):
    L = D.pop("target")  # Extract label L
    Dt = []

    for i in range(len(D.columns) - 2, -1, -1):
        feature = D.iloc[:, i]
        if feature.name not in Dt and feature.nunique() >= 20:
            transformed_feature = feature.apply(lambda x: L.iloc[int(x / 0.1)] if int(x / 0.1) < len(L) else x)
            Dt.append(transformed_feature)
    Dt = pd.DataFrame(Dt).T  # Convert list of series to DataFrame
  # Build a Random Forest model
    X_train, X_test, y_train, y_test = train_test_split(Dt, L, test_size=0.2, random_state=42)
    rf = RandomForestClassifier(max_depth=25, random_state=42)
    rf.fit(X_train, y_train)
    # Get feature importances
    d = dict(zip(Dt.columns, rf.feature_importances_))
    d_sorted = sorted(d.items(), key=lambda item: item[1], reverse=True
  # Select top n features
    Dp = Dt[[feature for feature, importance in d_sorted[:n]]]
    Dp["target"] = L  # Add label back
    return Dp
```

Algorithm-3 to perform PCA

```
# Sample dataset
np.random.seed(42)
# Standardize numerical features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(data[['feature1', 'feature2', 'feature3']])
# Perform PCA
pca = PCA(n_components=2)
principal_components = pca.fit_transform(X_scaled)
pc_df = pd.DataFrame(principal_components, columns=['PC1', 'PC2'])
# Add categorical variable
pc_df['category'] = data['category']
# Function to calculate correlation ratio (η²)
```

```
def correlation_ratio(cat, values):
    categories = np.unique(cat)
    total_variance = np.var(values)
    between_group_variance = sum(
        np.var(values[cat == c]) * len(values[cat == c]) for c in categories
    ) / len(values)
    return (total_variance - between_group_variance) / total_variance
# Compute η² for each principal component
correlations = {}
for pc in ['PC1', 'PC2']:
    correlations[pc] = correlation_ratio(pc_df['category'], pc_df[pc])

# Display correlation results
correlations_df = pd.DataFrame.from_dict(correlations, orient='index', columns=['Eta Squared'])
print(correlations_df)
```

2.1. Flowchart of the heart disease prediction.

Displays the distributions for Cleveland, Hungarian, and Long Beach, Virginia. Following the removal of any missing or worthless values, they have 280, 249, and 121 samples, respectively. Whereas the presence indicates illness, the absence indicates health.

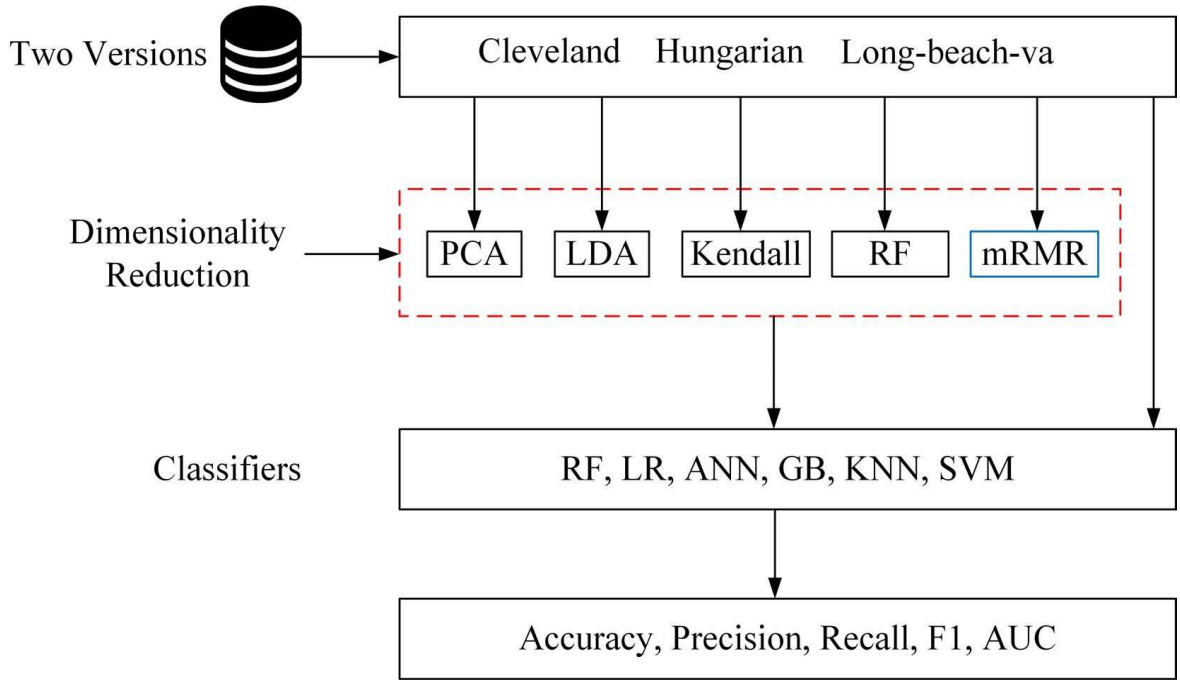

Fig1. Flowchart of heart disease prediction.

The flow chart describes that the heart disease prediction in various versins with the dimensionality reduction and classifiers.

3. OUTCOMES AND EXAMINATION

3.1. OUTCOMES FROM THE ACCIDENTAL DATASET

Both the datasets and the classifiers' properties have an impact on the prediction outcomes.With their respective AUCs, ANN and KNN exhibit the lowest performance among these classifiers, almost losing the prediction function. It suggests in an indirect way that pre-processing is required when using ANN and KNN, and that certain classifiers have a great ability to remove irrelevant information. The fact that dataset type affects the prediction effect is another noteworthy factor; Cleveland and Hungarian appear to be superior to Long Beach, Virginia.

The Kendell τ & the Random Forest methods, we evaluate the performance of the incremental feature combination method on three datasets in Fig. 2. In this experiment, 30 features are chosen and collected in a certain order. The first feature added is the 25th most important feature, followed by the 24th most important feature, and so on. The first features have a minimum of five. Based on the results, the accuracy of Cleveland and Hungarian increases quickly as the amount of features increases. On the Long-Beach-Va dataset, accuracy does not appear to increase until the feature number exceeds 25. The accuracy growth is modest and steady, particularly when the feature number is less than 15. It suggests that inconsequential characteristics have minimal impact.

Table 1: processed data sets features

S.No	Original number	Feature name	Feature description
1	3	Age	Age in years
2	4	Sex	Sex (1=male,0=female)
3	9	cp	Chest pain type • value1:typical angina, • value 2: a typical angina, • value 3: non anginal pain • value 4: asymptomatic
4	10	thresbps	The Resting hold presure
5	12	chol	The Serum cholestroal in mg/dl
6	16	fbs	The Fasting blood sugar>120mg/dl for 1=true,0=false
7	19	resteeg	The Resting electrocardiographic results
8	32	thalach	The Maximum heart rate achieved
9	38	exang	The Exercise induced angina for 1= yes, 0=no
10	40	oldpeak	The ST depression induced by exercise relative to rest
11	41	slope	The slope of the peak exercise ST segment
12	44	ca	Height at rest

13	51	thal	• 3=normal • 6=fixed defect • 7=reversable defect
14	58	num	Diagnosis of heart disease i.e. heart disease status.

4.Simulation Results

The simulation results for the proposed algorithms are as follows in the specified data sets.

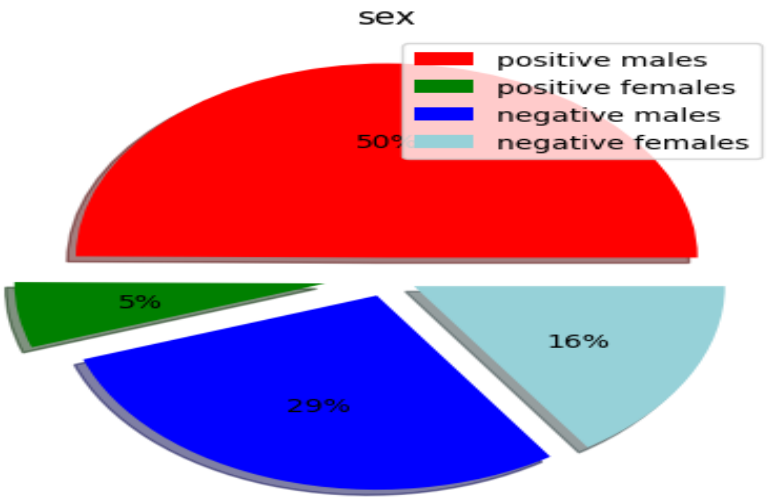

Fig2. distribution based on sex

In Fig 2,represents the distribution of test results based on sex, divided into four categories:

1. **Positive Males (Red - 50%):**This group makes up half of the total population, meaning that 50% of all individuals in the study are males who tested positive.

2. **Positive Females (Green - 5%):**This category is much smaller, showing that only 5% of the total individuals are females who tested positive.

3. **Negative Males (Blue - 29%):**This segment represents 29% of the total population, indicating males who tested negative.

4. **Negative Females (Light Blue - 16%):**This section accounts for 16% of the total population, representing females who tested negative.

The given pie chart represents the relationship between blood sugar levels and heart disease. It is divided into two segments, showing the percentage of individuals with heart disease who either have high blood sugar or do not.Dark Purple Section (33%)- represents people who have high blood sugar and also suffer from heart disease. Light Blue Section (67%) -represents people who do not have high blood sugar but still suffer from heart disease.

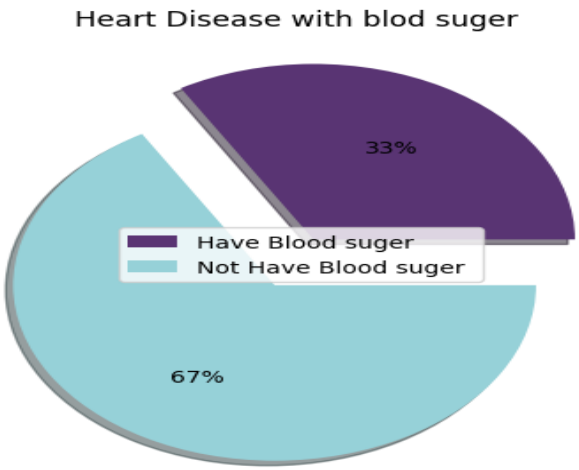

Fig 3. distribution of blood sugar levels

A majority (67%) do not have high blood sugar**.** However**,** 33% of people with heart disease high blood sugar**,** indicating a possible link between blood sugar levels and heart disease. The 3D effect and the separation of one **slice** (exploded pie chart) are likely used to emphasize the Detailed Explanation of the Pie Chart on Heart Disease and Blood Sugar. In Fig 3, represents the distribution of blood sugar levels among individuals with heart disease**.**

The Fig 3, is divided into two portions:

- 33% (Dark Purple Segment): These are individuals who have high blood sugar and also suffer from heart disease**.**
- 67% (Light Blue Segment): These are individuals who do not have high blood sugar but still suffer from heart disease

Blood Sugar Risk Factor:

High blood suar (diabetes or prediabetes) is a well-known risk factor for heart .The 33% figure suggests that a significant proportion of heart disease cases are associated with high blood sugar, possibly due to conditions like diabetes, insulin resistance, or metabolic syndrome .Other Causes of Heart Disease (67% Group): The 67% who do not have high blood sugar still have heart disease, meaning there are other risk factors contributing to their condition,

- High blood pressure
- Smoking
- High cholesterol
- Obesity
- Lack of exercise
- Genetics (family history of heart disease)

- **Clinical Importance:**
- While diabetes is a major contributor to heart disease, it is not the only cause**.**
- Even people with normal blood sugar levels should maintain a heart-healthy lifestyle to reduce their risk.

Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 46 No. 2 (2025)

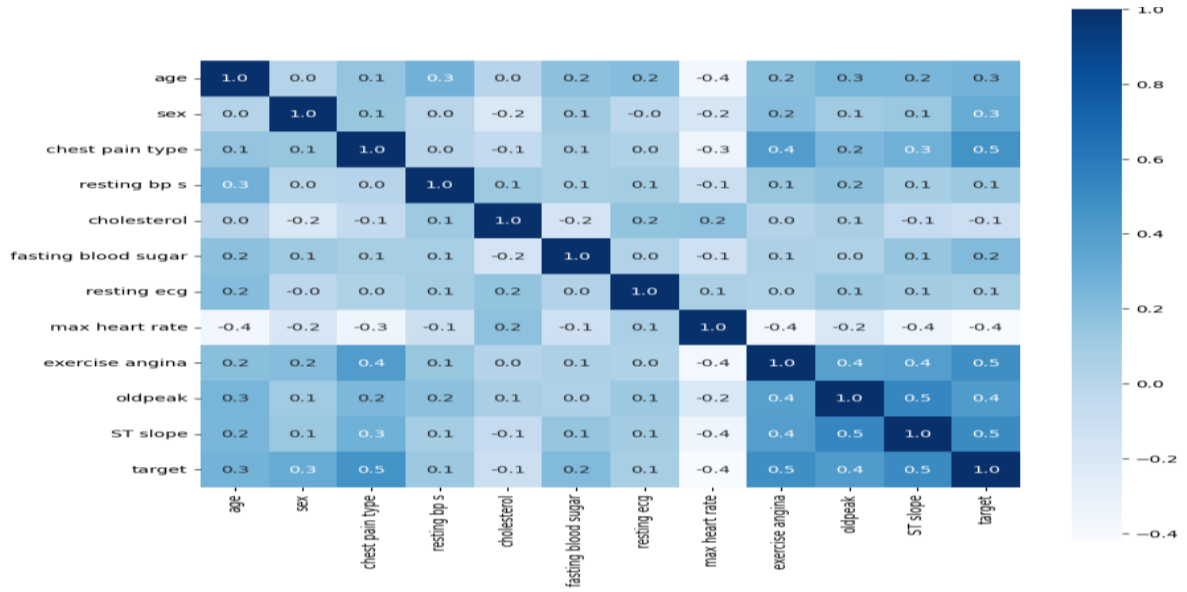

Fig 4. Heatmap of features of the Cleveland CHD dataset derived using Pearson correlation coefficient

In Fig 4, represent Heatmap of features of the Cleveland CHD dataset derived using Pearson correlation coefficient dataset, derived using Pearson correlation coefficient. The Pearson correlation measures the linear dependency between two variables, with values ranging from -1 to 1. In the heatmap, red represents negatively correlated features, meaning an increase in one feature leads to a decrease in the other, while green indicates positively correlated features, signifying that both variables increase together. This visual representation helps identify patterns that may contribute to heart disease prediction.

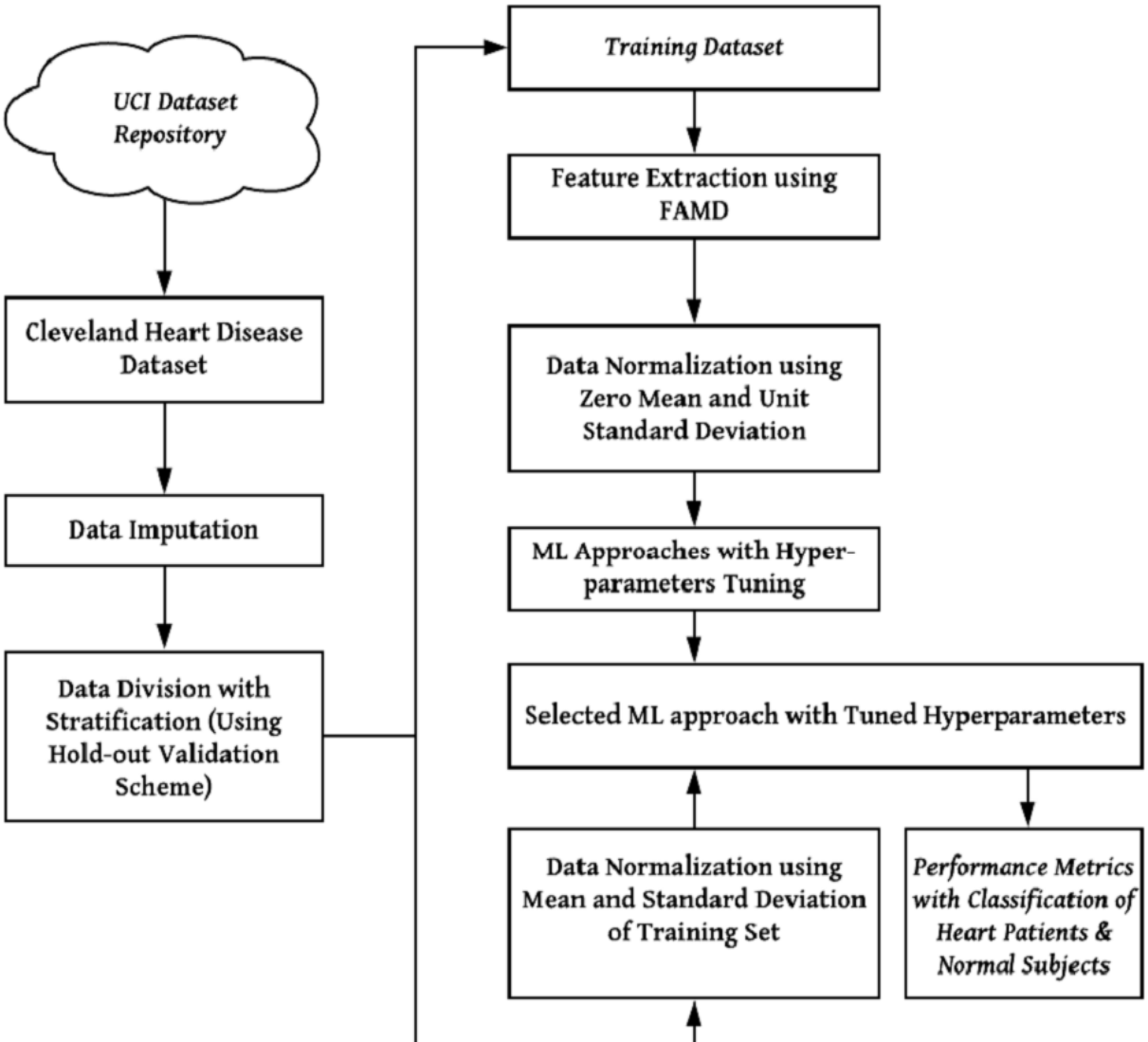

Fig 5. A machine intelligence framework for Cleveland heart disease dataset.

.**Table 2. Proposed different aspect and details**

Aspect	**Details**
Correlation Method	Pearson Correlation Coefficient
Heatmap Color Scale	- Red: Negatively correlated features - Green: Positively correlated features
Correlation Range	-1 (strong negative) to +1 (strong positive)
Dataset Name	Cleveland CHD Dataset
Key Features	Age, Sex, Cholesterol, Resting Blood Pressure, ECG Readings, etc.
Total Instances	Includes both normal subjects and heart patients
Data Representation	Graphical representation of dataset composition
Imputation	Used to handle missing values for complete analysis
Purpose	Identify key predictors of heart disease, assist in early diagnosis and treatment

By analysing the correlation heatmap, researchers can determine which features have the most significant impact on heart disease risk. Strong correlations may indicate potential

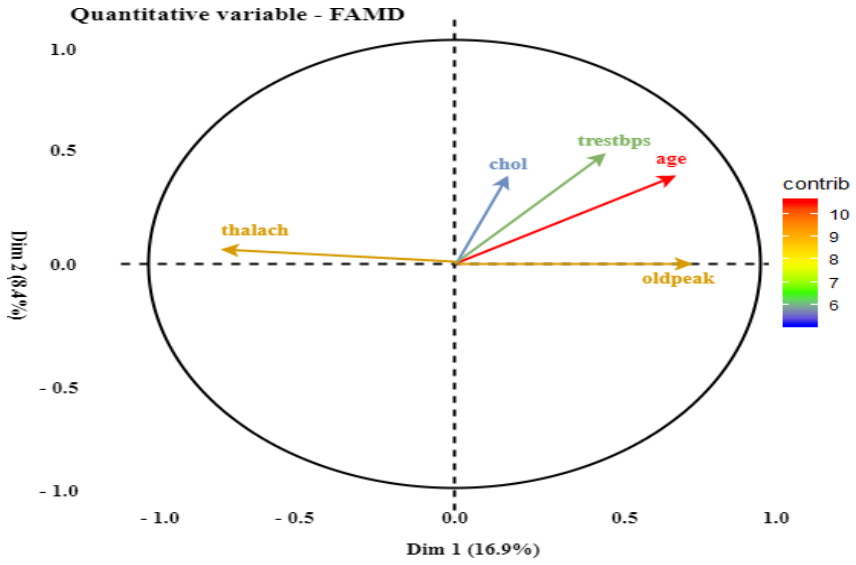

Fig 6 correlation between categorical features & dimensions

biomarkers or key predictors that can aid in early diagnosis and targeted treatment strategies

In Fig7, represents this is a confusion matrix for an SVM classification. **True Positives (TP)**: 95 (Actual Positive, Predicted Positive)

- **False Negatives (FN)**: 23 (Actual Positive, Predicted Negative)
- **False Positives (FP)**: 20 (Actual Negative, Predicted Positive)
- **True Negatives (TN)**: 138 (Actual Negative, Predicted Negative)

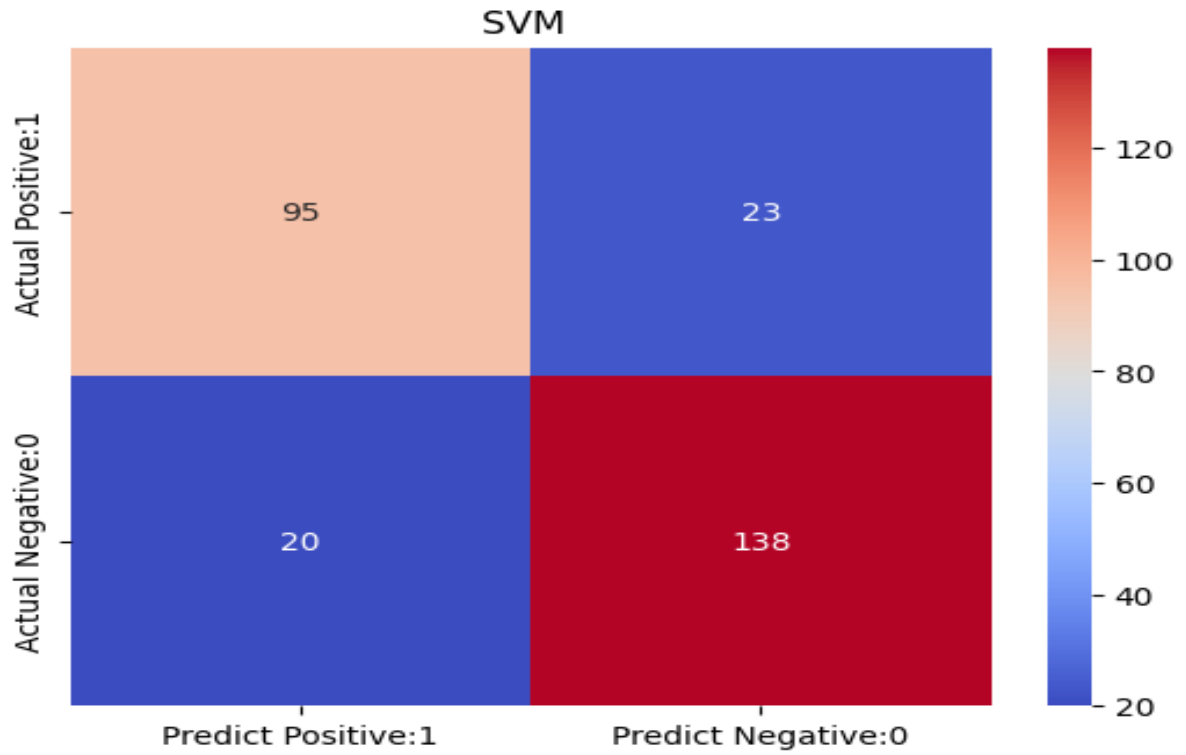

Fig 7. confusion matrix for an SVM

From this, you can calculate key performance metrics:

- **Accuracy** = (TP + TN) / (TP + TN + FP + FN) = (95 + 138) / (95 + 138 + 20 + 23) = 233 / 276 ≈ **84.42%**
- **Precision (Positive Predictive Value)** = TP / (TP + FP) = 95 / (95 + 20) = 95 / 115 ≈ **82.61%**
- **Recall (Sensitivity, True Positive Rate)** = TP / (TP + FN) = 95 / (95 + 23) = 95 / 118 ≈ **80.51%**
- **Specificity (True Negative Rate)** = TN / (TN + FP) = 138 / (138 + 20) = 138 / 158 ≈ **87.34%**
- **F1 Score** = 2 × (Precision × Recall) / (Precision + Recall) ≈ 2 × (0.8261 × 0.8051) / (0.8261 + 0.8051) ≈ **81.75%**

This shows that the SVM model performs well, with a good balance between precision and recall. If you want improvements, you might fine-tune hyperparameters or try different kernel functions.

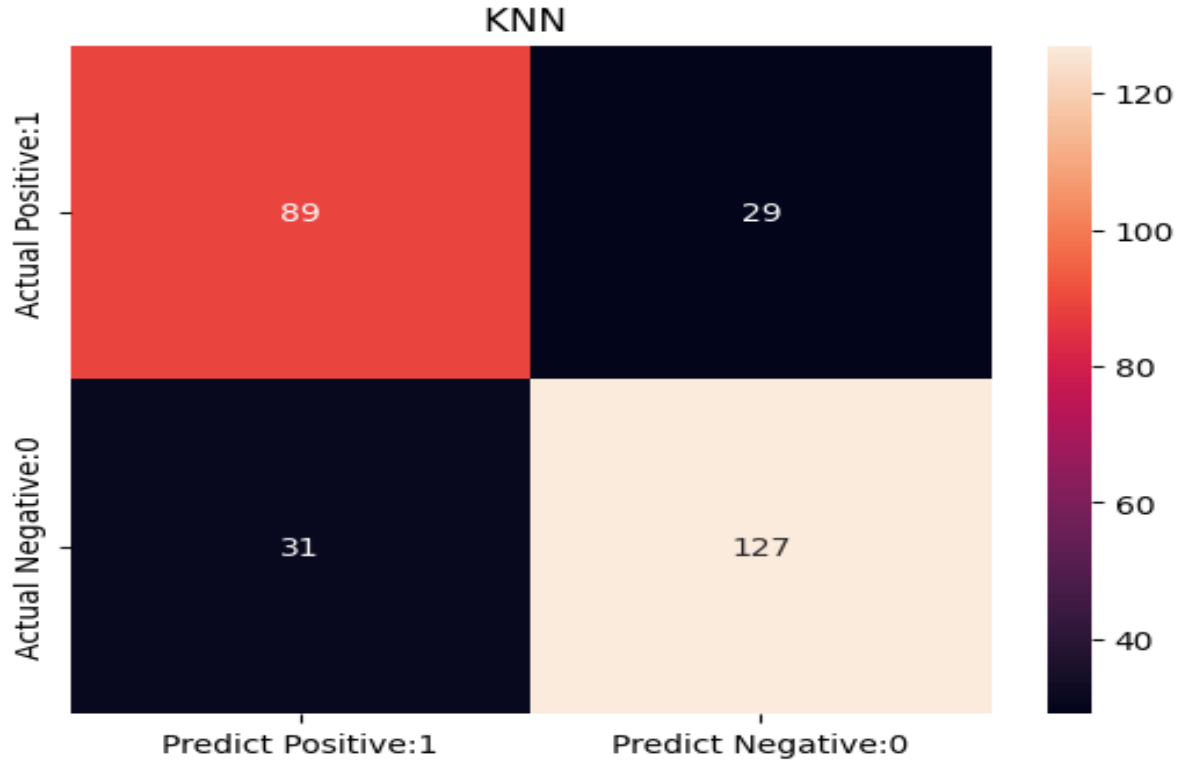

Fig 8. confusion matrix for a K-Nearest Neighbours (KNN)

This is a confusion matrix for a K-Nearest Neighbours (KNN) classification model. Here's the breakdown of the values:

- **True Positives (TP)**: 89 (correctly predicted positive cases)
- **False Negatives (FN)**: 29 (actual positives incorrectly classified as negatives)
- **False Positives (FP)**: 31 (actual negatives incorrectly classified as positives)
- **True Negatives (TN)**: 127 (correctly predicted negative cases)

From this, you can calculate key performance metrics:

1. **Accuracy** = $\frac{TP+TN}{TP+TN+FP+FN} = \frac{89+127}{89+127+31+29} = \frac{216}{276} = 0.783 = 78.3\%$
2. **Precision** = $\frac{TP}{TP+FP} = \frac{89}{89+31} = \frac{89}{120} = 0.742 = 74.2\%$
3. **Recall (Sensitivity)** = $\frac{TP}{TP+FN} = \frac{89}{89+29} = \frac{89}{118} = 0.754 = 75.4\%$
4. **F1-Score** = $2 \times \frac{Precision \times Recall}{Precision+Recall} = 0.748 = 74.8$

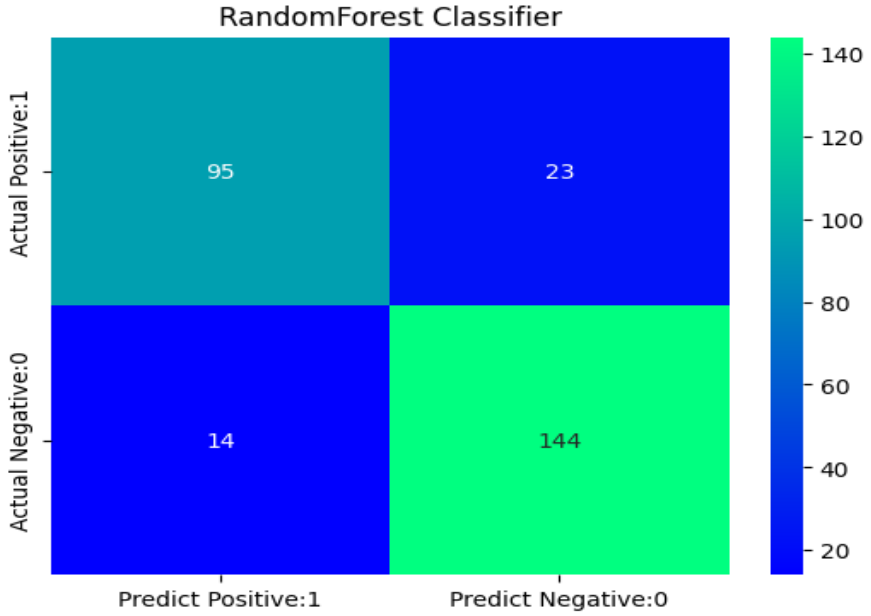

Fig 9. Random Forest Classifier

In Fig 9, represents:

- **True Positives (TP):** 95 (Actual Positive, Predicted Positive)
- **False Positives (FP):** 14 (Actual Negative, Predicted Positive)
- **False Negatives (FN):** 23 (Actual Positive, Predicted Negative)
- **True Negatives (TN):** 144 (Actual Negative, Predicted Negative)

This confusion matrix represents the performance of a **Random Forest Classifier** in a binary classification problem. It compares the actual labels with the predicted labels to assess **True Positives (TP) = 95**

- The model correctly identified 95 positive cases.
- **False Negatives (FN) = 23**
- The model incorrectly predicted 23 actual positive cases as negative.
- **False Positives (FP) = 14**
- The model incorrectly predicted 14 actual negative cases as positive.
- **True Negatives (TN) = 144**
- The model correctly identified 144 negative cases.

Using these values, we can calculate the following key metrics:

- **Accuracy:** 86.6% (Overall correctness)
- **Precision:** 87.2% (Reliability of positive predictions)
- **Recall:** 80.5% (Ability to detect actual positives)
- **F1-Score:** 83.7% (Harmonic mean of precision & recall)

The model performs well overall**,** with a high precision (87.2%)**,** meaning it makes few false positive errors.

Table3. Performance proposed method MIFH along with baseline methods statistics on UCI heart disease Cleveland.

Research by	Method	Accuracy in %	Sensitivity in %	Specificity in %
Purushottam et all in 2016	Rule based classifier	86.7	…	…
Sha etc all in 2017	PPCA	82.18	75	90.5
Vijayashree etc all in 2018	PSO with SVM	84.36	---	---
Haq etc all in 2018	Relief +LR	89	77	98
Haq etc all in 2018	mRMR+NB	84	77	90
Hq etc all in 2019	LASSO+SVM	8	75	96
Saqline etc all in 2019	RBF kernel based SVM	81.19	72.92	88.68
Mohan etc all in 2019	HRF LM	88.7	92.8	82.6
Ali etc all in 2019	L1 linear SVM + L2 linear & RBF SVM	92.22	82.92	100
MIFH	FAMD +RF	93.44	89.28	96.96

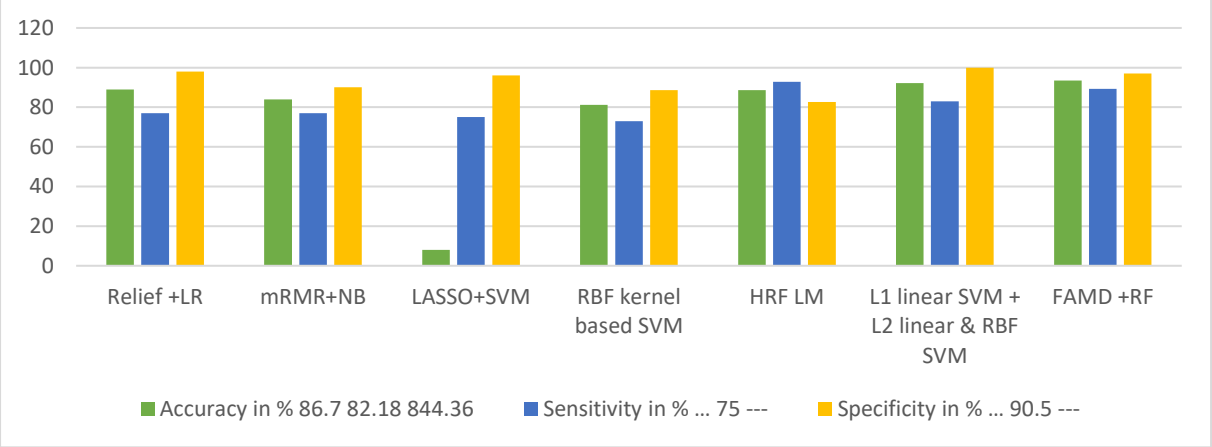

Fig 10. *MIFH* along with baseline methods statistics on UCI heart disease Cleveland dataset.

Table 4. report of SVM, KNN etc

Report of SVM

	precision	recall	f1-score	support
0	0.83	0.81	0.82	118
1	0.86	0.87	0.87	158

accuracy			0.84	276
macro avg	0.84	0.84	0.84	276
weighted avg	0.84	0.84	0.84	276

##

Report of KNN

	precision	recall	f1-score	support
0	0.74	0.75	0.75	118
1	0.81	0.80	0.81	158
accuracy			0.78	276
macro avg	0.78	0.78	0.78	276
weighted avg	0.78	0.78	0.78	276

##

Report of SVM

	precision	recall	f1-score	support
0	0.87	0.81	0.84	118
1	0.86	0.91	0.89	158
accuracy			0.87	276
macro avg	0.87	0.86	0.86	276
weighted avg	0.87	0.87	0.87	276

##

Report of KNN

	precision	recall	f1-score	support
0	0.74	0.75	0.75	118
1	0.81	0.80	0.81	158
accuracy			0.78	276

macro avg	0.78	0.78	0.78	276
weighted avg	0.78	0.78	0.78	276

##

Report of SVM

	precision	recall	f1-score	support
0	0.87	0.81	0.84	118
1	0.86	0.91	0.89	158
accuracy			0.87	276
macro avg	0.87	0.86	0.86	276
weighted avg	0.87	0.87	0.87	276

##

Report of SVM

	precision	recall	f1-score	support
0	0.80	0.81	0.80	118
1	0.85	0.85	0.85	158
accuracy			0.83	276
macro avg	0.83	0.83	0.83	276
weighted avg	0.83	0.83	0.83	276

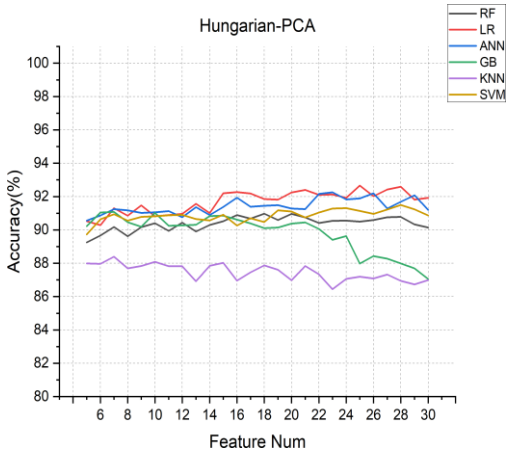

Fig 11. Hungarian-PCA.

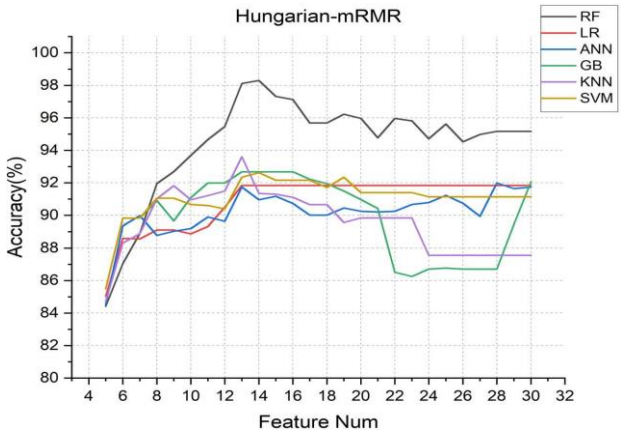

Fig 12. Hungarian mRMR

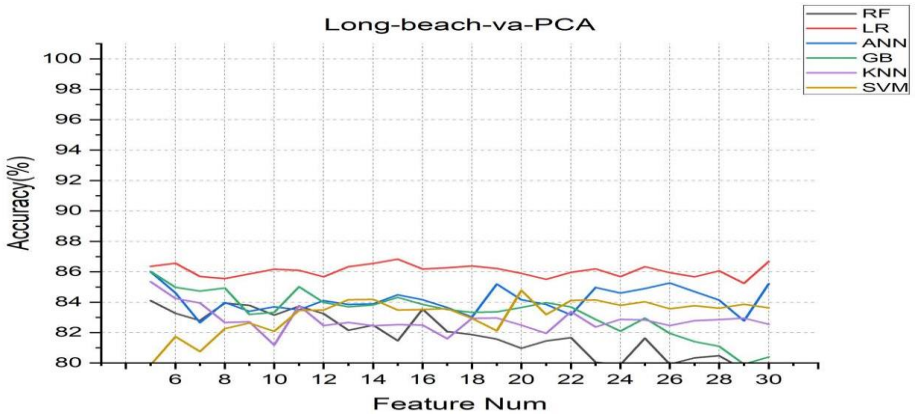

Fig 13.Long bech-va-PCA

Table 5. Data set of Long-Beach-Va

Data set	Method	Accuracy (%)		Precision (%)		F 1(%)		Recall (%)	
Long-Beach-Va	RF	76.2	87.3	79.7	87.7	85.2	92.1	91.8	97.1
	LR	74.0	82.5	75.7	88.3	84.4	88.4	95.8	88.7
	ANN	71.4	68.5	73.4	77.1	82.8	79.6	95.5	82.8
	GB	68.8	92.6	82.1	92.6	78.1	95.0	75.4	97.6
	KNN	73.2	73.0	76.7	76.1	83.7	83.7	92.7	93.4
	SVM	64.5	74.9	84.6	87.0	82.0	82.0	63.6	77.8

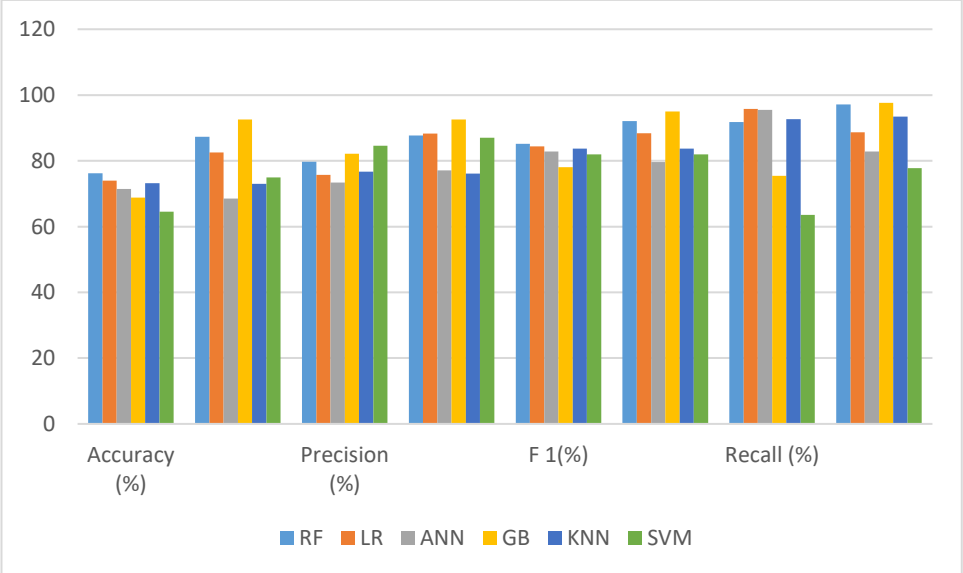

Fig 14. Comparision of Long bech-va-PCA

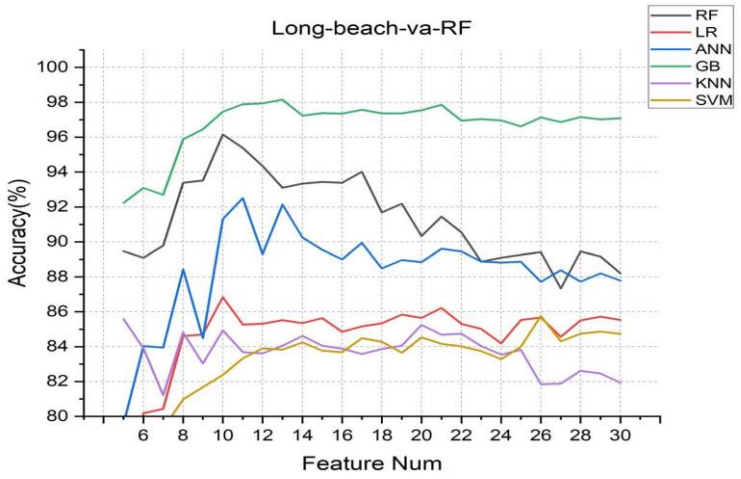

Fig15. Long beach-va-R

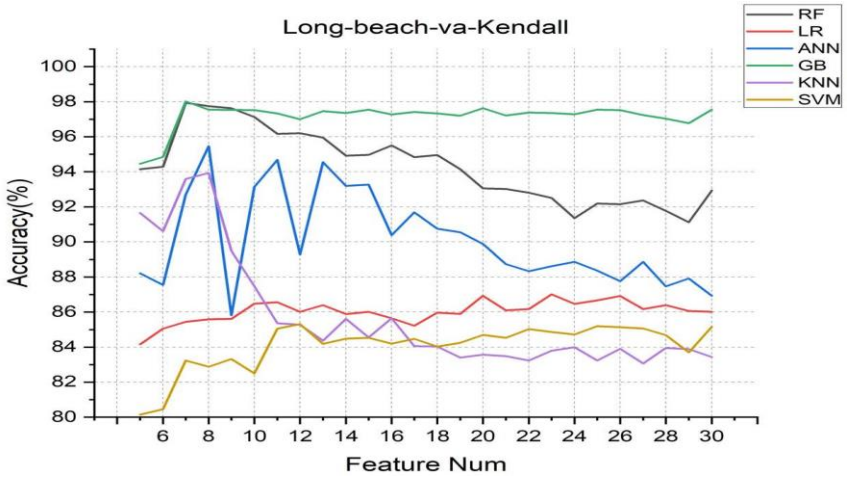

Fig16. Long beach-va -Kendall

ss

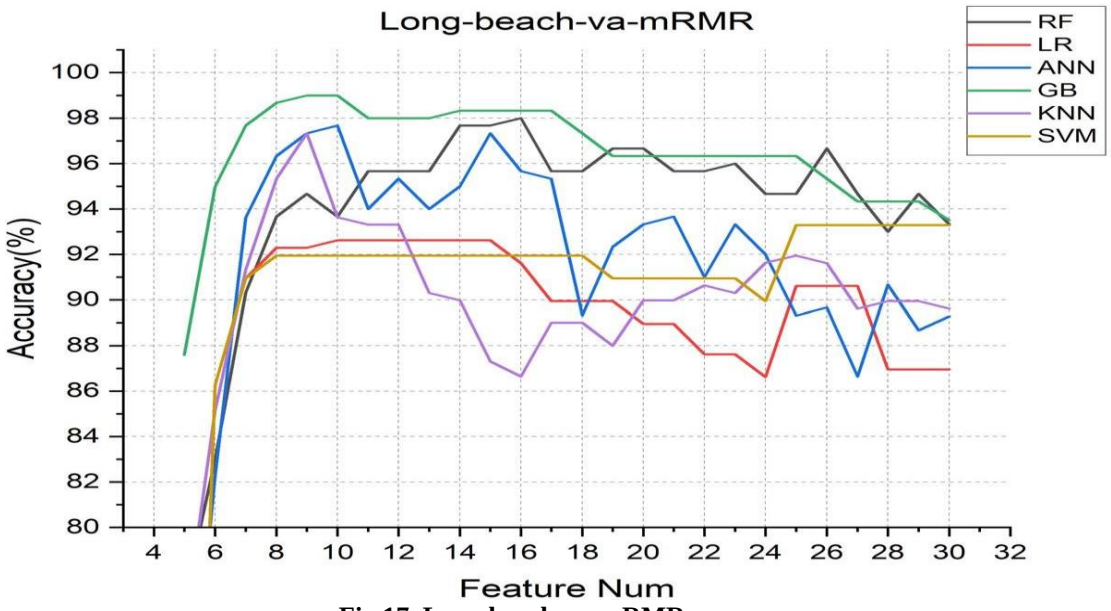

Fig 17. Long beach-va-mRMR

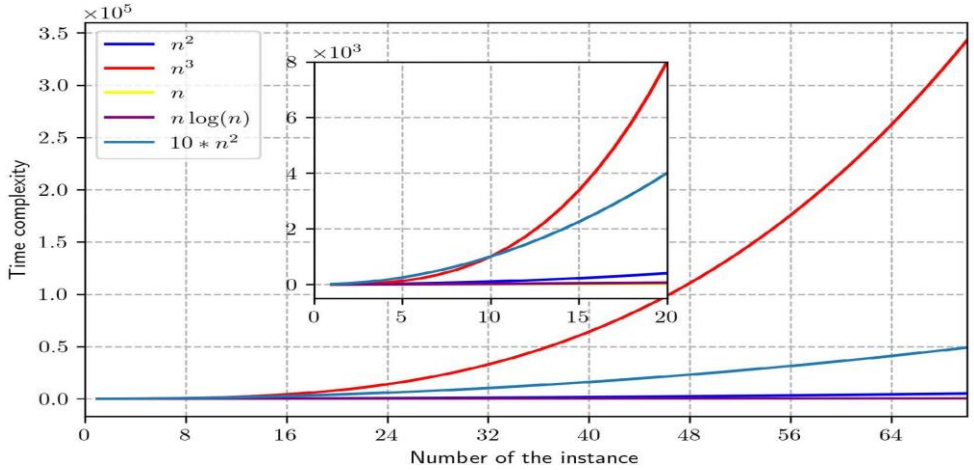

Fig 18.Time complexity with the instances.

Table 6. Comparision on Cleveland data set

Data set	Method	Accuracy (%)		Precision (%)		F 1(%)		Recall (%)	
Cleveland	RF	81.5	94.6	81.2	97.2	78.0	93.5	75.7	90.2
	LR	83.2	92.6	86.6	97.4	81.4	90.9	77.2	85.5
	ANN	61.4	57.1	58.2	46.8	59.4	49.6	68.2	62.0
	GB	83.7	98.8	84.8	97.8	81.6	98.6	79.1	99.4
	KNN	63.8	63.1	59.4	60.9	58.4	54.5	57.9	49.9
	SVM	85.1	89.5	85.5	91.0	83.6	87.2	82.1	84.1

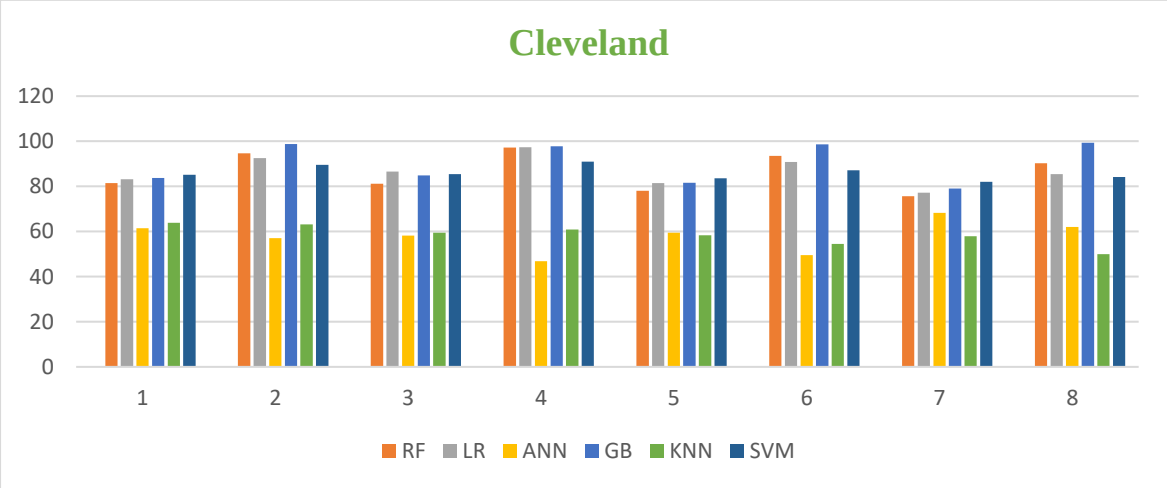

Fig 19. Comparision on Cleveland data set

Table 7. Comparision on Hungarian data set

	Method	Accuracy (%)		Precision (%)		F 1(%)		Recall (%)	
	RF	80.9	92.6	75.9	91.5	72.2	89.7	69.6	88.2

Hungarian	**LR**	83.0	91.0	78.4	88.6	73.4	87.1	69.6	86.2
	ANN	63.4	63.3	11.3	3.5	3.1	1.3	15.6	0.8
	GB	82.6	84.8	76.4	76.5	77.7	79.8	79.5	84.8
	KNN	64.0	68.4	52.5	58.4	44.1	45.6	38.8	39.6
	SVM	80.9	89.6	73.4	85.3	72.7	85.5	72.7	86.1

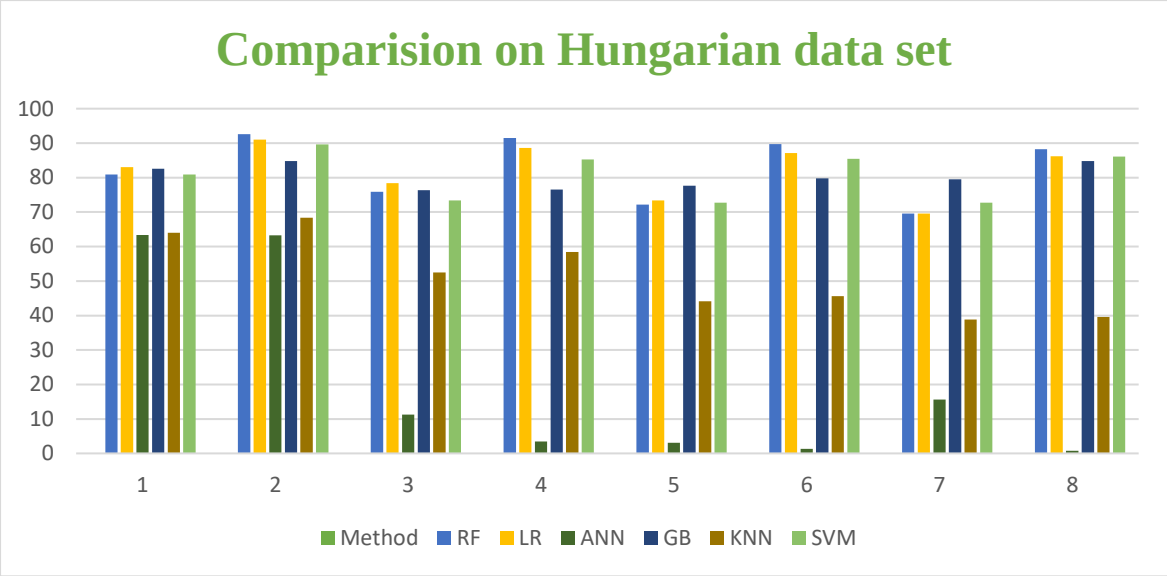

Fig 20.Comparision on Hungarian data set

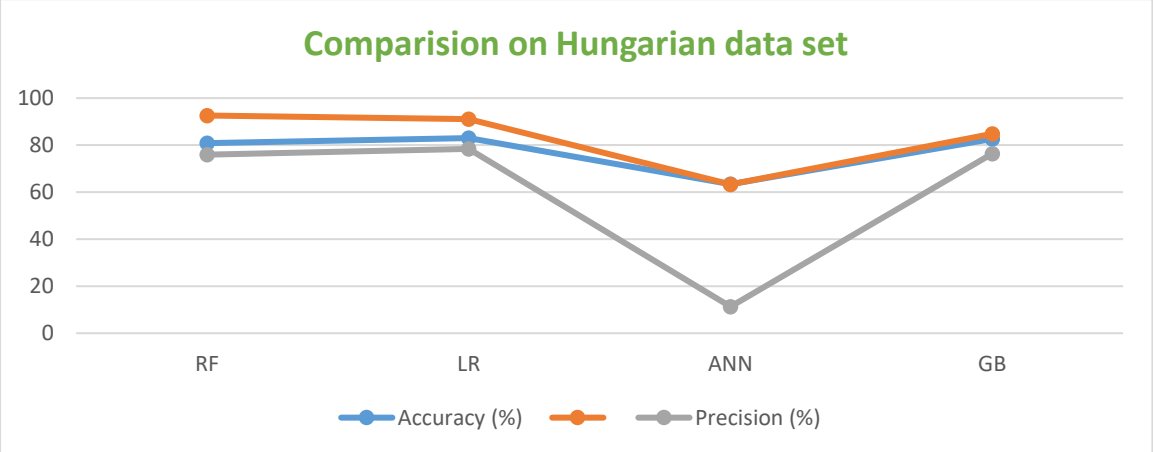

Fig 21.Comparision on Hungarian data set

Table 8.Method, accuracy and number of features details

Author	**Method**	**Accuracy**	**Feature**
D. Khanna, R. Sahu	Logistic Regression	**84.85%**	13
S. S. Khan and S. Quadri	Decision tree	**89.1%**	13
D. R. V. S. Kodati	Navie bayes	**83.7%**	13

F. S. Alotaibi	SVM	**92.3%**	13
K. Uyar and A. Ilhan	GF based on RFNN	**96.63%**	13
A. Gupta, & etc	MIFH	**93.44%**	28
Proposed	mRMr+RF, ANN, SVM	**97.3%**	13

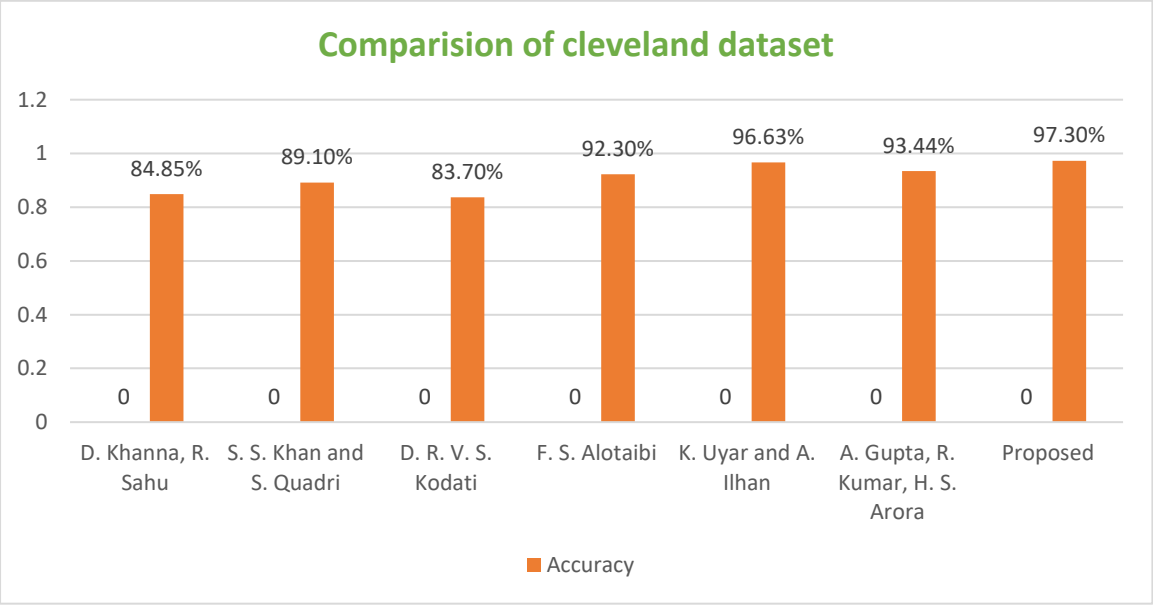

Fig 22. Comparision of Cleveland data set

Table 9.Selected features by Kendall from original data:

Method		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Kendall	**Cleveland**	61	67	51	65	63	60	44	38	9	68	40	41	4
	Hungaria	63	38	60	41	6	40	67	9	30	7	72	65	68
	Long-beach-va	67	63	60	6	9	61	72	38	40	41	7	68	39

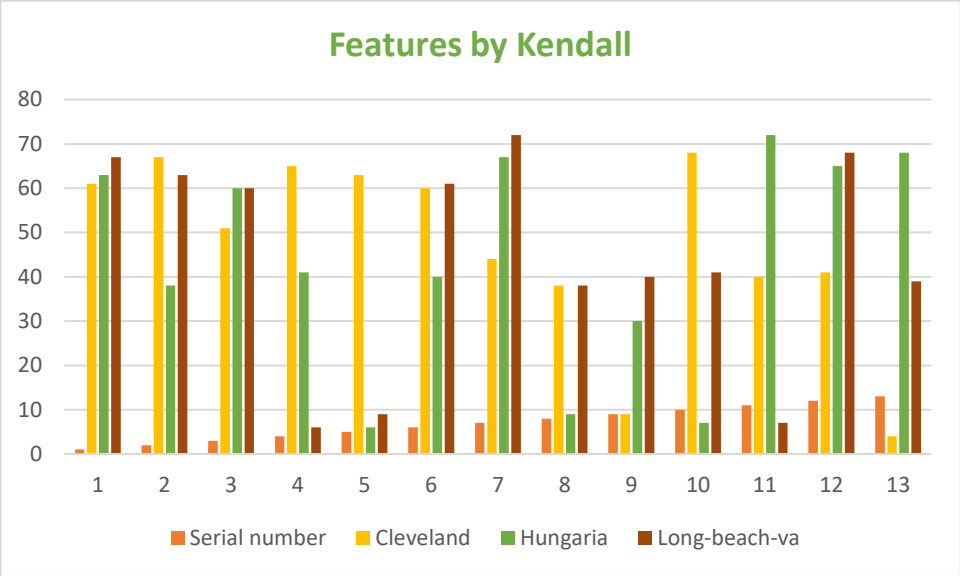

Fig 23. Features by Kendall from original data sets

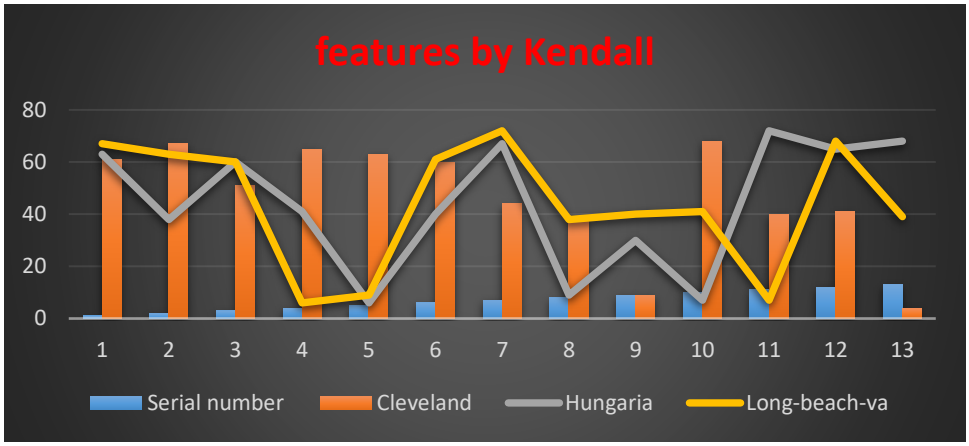

Fig 24. Features by Kendall from original data sets-1

Table 10. Selected features by Kendall from original data sets

Method	**Serial number**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Random forest	**Cleveland**	51	61	63	9	60	67	38	40	65	32	44	30	68
	Hungaria	63	38	41	9	6	60	7	61	67	40	65	32	30
	Long-beach-va	67	60	63	3	12	40	56	61	32	65	6	74	29

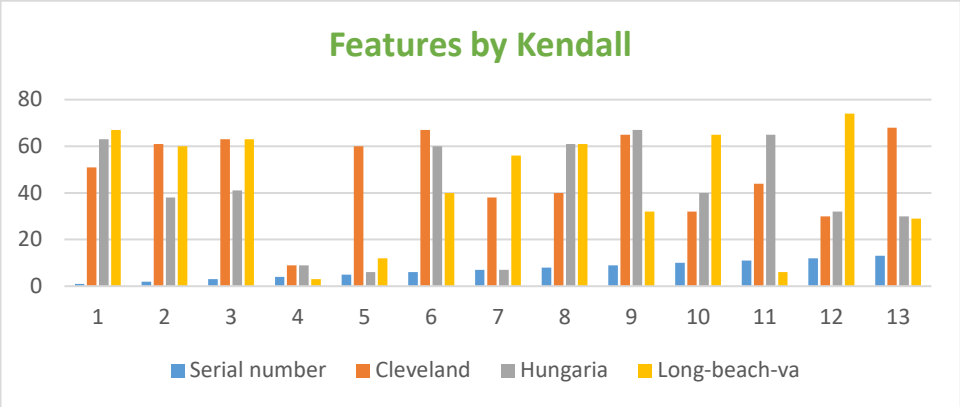

Fig 24. Features by Kendall from original data sets-2

Table 11.Selected features with mRMR from original data sets

Method	Serial number	1	2	3	4	5	6	7	8	9	10	11	12	13
mRMR	Cleveland	5	42	41	45	43	44	40	46	35	29	34	2	39
	Hungaria	9	58	45	48	38	52	3	46	35	53	50	55	44
	Long-beach-va	9	5	58	56	52	49	50	54	57	3	48	53	1

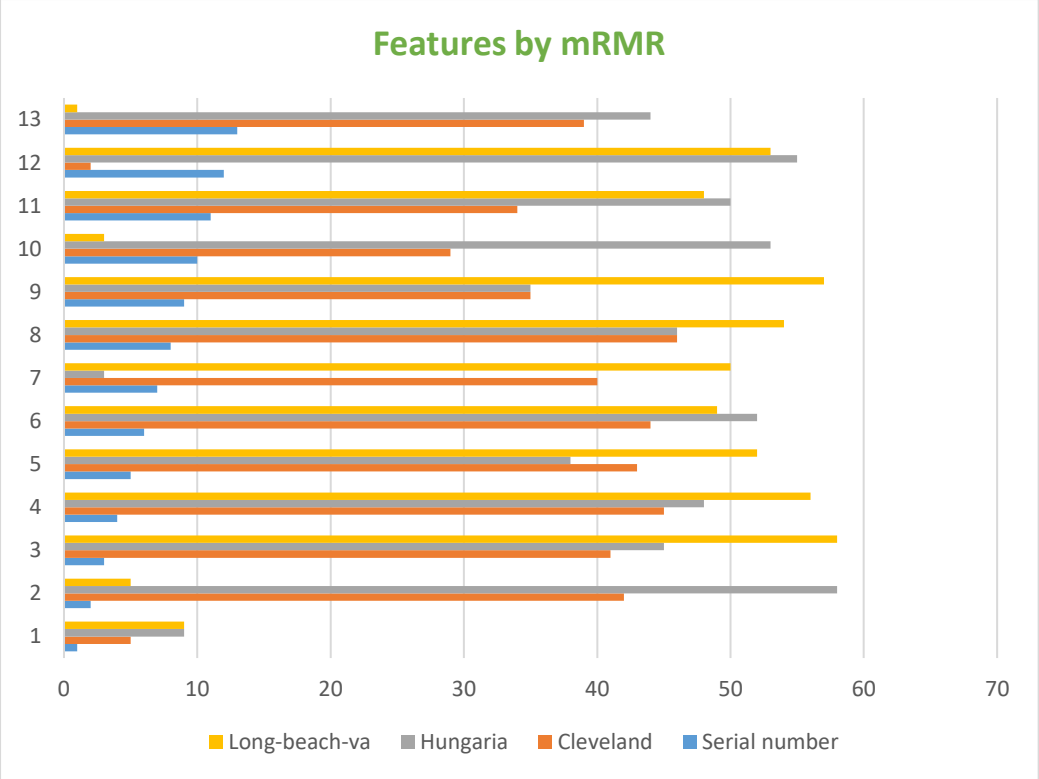

Fig 23. Features by mRMR from original data sets

5. Conclusion

This research demonstrates that mRMR is an effective feature selection method for heart disease prediction. Our findings show that reducing features indiscriminately degrades classifier performance, while selecting the most relevant ones enhances accuracy. The incremental feature selection method achieves over 90% of the best performance, highlighting the importance of key features.

mRMR outperforms other selection methods by consistently improving classifier accuracy across datasets. Among the tested classifiers, Random Forest, Gaussian Bayesian, and ANN exhibit the best predictive performance. These results confirm the value of mRMR in optimizing feature selection for heart disease prediction. Additionally, eliminating irrelevant or redundant features not only improves model performance but also enhances computational efficiency, reducing training time and overfitting risks. These findings emphasize the importance of feature selection in medical diagnostics and suggest that mRMR can be a valuable tool for improving predictive modelling in healthcare applications.

6. REFERENCES

1. A. S. Eltrass, "Novel cascade filter design of improved sparse low-rank matrix estimation and kernel adaptive filtering for ECG denoising and artifacts cancellation," *Biomed. Signal. Process. Control.*, Vol. 77, pp. 103750, 2022.
2. B.Tutuko,*etal*.,"DAE-ConvBiLSTM:end-to-endlearning single-lead electrocardiogram signal for heart abnormal- ities detection," *PLoS One*, Vol. 17, no. 12, pp. e0277932, 2022.
3. A. Kumar, M. Kumar, and R. S. Komaragiri, "ECG sig- nal denoising techniques for cardiac pacemaker systems," in *High performance and power efficient electrocardiogram detectors*, Singapore: Springer Nature Singapore, 2022, pp. 49–78.
4. S. Kiranyaz, *et al*., "Blind ECG restoration by operational cycle-GANs," *IEEE Trans. Biomed. Eng.*, Vol. 69, no. 12, pp. 3572–81, 2022.
5. V. Patel, and A. Shah, "Denoising electrocardiogram sig- nals using multiband filter and its implementation on FPGA," *Serbian Journal of Electrical Engineering*, Vol. 19, no. 2, pp. 115–28, 2022.
6. S. Murawwat, *et al*., "Denoising and classification of arrhythmia using memd and ann," *Alexandria Eng. J.*, Vol. 61, no. 4, pp. 2807–23, 2022.
7. A. S. Eltrass, "Novel cascade filter design of improved sparse low-rank matrix estimation and kernel adaptive filtering for ECG denoising and artifacts cancellation," *Biomed. Signal. Process. Control.*, Vol. 77, pp. 103750, 2022.
8. A. Abdallah, *et al*. "ECG signal denoising based on wavelet transform and genetic algorithm," in *2023 International Conference on Advances in Electronics, Control and Com- munication Systems (ICAECCS)*. IEEE, 2023.
9. J. D. K. Abel, S. Dhanalakshmi, and R. Kumar, "A com- prehensive survey on signal processing and machine learn- ing techniques for non-invasive fetalecg extraction," *Mul- timed. Tools. Appl.*, Vol. 82, no. 1, pp. 1373–400, 2023.
10. M. R. Laskar, *et al*. "A complexity efficient penta-diagonal quantum smoothing filter for ECG signal denoising," (2023).
11. A. Rasti-Meymandi, and A. Ghaffari, "A deep learning- based framework For ECG signal denoising based on stacked cardiac cycle tensor," *Biomed. Signal. Process. Con- trol.*, Vol. 71, pp. 103275, 2022.
12. X. Wang, *et al*., "An ECG signal denoising method using conditional generative adversarial net," *IEEE. J. Biomed. Health. Inform.*, Vol. 26, no. 7, pp. 2929–40, 2022.
13. P.Madan,*etal*.,"DenoisingofECGsignalsusingweighted stationary wavelet total variation," *Biomed. Signal. Process. Control.*, Vol. 73, pp. 103478, 2022.

14. M. Das, and B. C. Sahana, "Optimized orthogonal wavelet-based filtering method for electrocardiogram sig- nal denoising," *J. Inst. Eng. (India): B.*, 1–14, 2022.

15. N. Li, *et al*., "The identification of ECG signals using WT- UKF and IPSO-SVM," *Sensors*, Vol. 22, no. 5, pp. 1962, 2022.

16. S. Balasubramanian, M. S. Naruk, and G. Tewari, "Elec- trocardiogram signal denoising using optimized adaptive hybrid filter with empirical wavelet transform," *J. Shanghai Jiaotong Univ. (Sci.)*, Vol. 30, no. 1, pp. 66–80, 2025.

17. H. Lin, R. Liu, and Z. Liu, "ECG signal denoising method based on disentangled autoencoder," *Electronics. (Basel)*, Vol. 12, no. 7, pp. 1606, 2023.

18. C. Chen, *et al*., "Wavelet-domain group-sparse denoising method for ECG signals," *Biomed. Signal. Process. Control.*, Vol. 83, pp. 104702, 2023.

19. J. Mao, *et al*., "A novel ECG signal denoising algorithm based on sparrow search algorithm for optimal variational modal decomposition," *Entropy* , Vol. 25, no. 5, pp. 775, 2023.

20. Y. Hou, *et al*., "Deep neural network denoising model based on sparse representation algorithm for ECG signal," *IEEE Trans. Instrum. Meas.*, Vol. 72, pp. 1–11, 2023.

21. P. M. Tripathi, *et al*., "A novel approach for real- time ECG signal denoising using Fourier decomposition method," *Res. Biomed. Eng.*, Vol. 38, no. 4, pp. 1037–49, 2022.

22. A. Mishra, *et al*., "ECG data analysis with denoising approach and customized CNNs," *Sensors*, Vol. 22, no. 5, pp. 1928, 2022.

23. L. Gao, Y. Gan, and J. Shi, "A novel intelligent denoising method of ecg signals based on wavelet adaptive threshold and mathematical morphology," *Appl. Intel.*, Vol. 52, no. 9, pp. 10270–84, 2022.

24. S. A. Malik, S. A. Parah, and B. A. Malik, "Power line noise and baseline wander removal from ECG signals using empirical mode decomposition and lifting wavelet trans- form technique," *Health. Technol. (Berl)*, Vol. 12, no. 4, pp. 745–56, 2022.

25. P. De Luca, A. Galletti, and L. Marcellino. "An accelerated algorithm for ECG signal denoising," in *2022 16th Inter- national Conference on Signal-Data Technology & Internet- Based Systems (SITIS)*. IEEE, 2022.

26. P. Upadhyay, S. K. Upadhyay, and K. K. Shukla, "Schrödinger equation based ECG signal denoising," *Chin. J. Phys.*, Vol. 77, pp. 2238–57, 2022.

27. H. Yang, and Z. Wei, "An effective morphological-stabled denoising method for ECG signals using wavelet-based techniques," *Int. J. Biomed. Eng. Technol.*, Vol. 39, no. 3, pp. 263–82, 2022.

28. P. Singh, and A. Sharma, "Attention-based convolutional denoising autoencoder for two-lead ECG denoising and arrhythmia classification," *IEEE Trans. Instrum. Meas.*, Vol. 71, pp. 1–10, 2022.

29. A. Paul, *et al*., "Automated detection of cardinal points of ECG signal for feature extraction using a single median filter," *J. Inst. Eng. (India): B.*, 1–12, 2022.

30. P. M. Tripathi, *et al*., "Watermarking of ECG signals compressed using Fourier decomposition method," *Mul- timed. Tools Appl.*, Vol. 81, no. 4, pp. 19543–19557, 2022.

31. D. Khanna, R. Sahu, V. Baths, and B. Deshpande, ''Comparative study of classification techniques (SVM, logistic regression and neural networks) to predict the prevalence of heart disease,'' *Int. J. Mach. Learn. Comput.*, vol. 5, no. 5, p. 414, 2015.

32. S. S. Khan and S. Quadri, ''Prediction of angiographic disease status using rule based data mining techniques,'' *Biol. Forum, Int. J.*, vol. 8, no. 2,pp. 103–107, 2016.

33. D. R. V. S. Kodati, ‘‘Analysis of heart disease using in data mining tools Orange and Weka,’’ *Global J. Comput. Sci. Technol.*, vol. 18, no. 1,pp. 17–21, Feb. 2018.

34. F. S. Alotaibi, ‘‘Implementation of machine learning model to predict heart failure disease,’’ *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 6, pp. 261–268,2019.

35. K. Uyar and A. Ilhan, ‘‘Diagnosis of heart disease using genetic algo- rithm based trained recurrent fuzzy neural networks,’’ *Proc. Comput. Sci.*, vol. 120, pp. 588–593, Jan. 2017.

36. A. Gupta, R. Kumar, H. S. Arora, and B. Raman, ‘‘MIFH: A machine intelligence framework for heart disease diagnosis,’’ *IEEE Access*, vol. 8,pp. 14659–14674, 2020.

Authors Profile:

Rajani Alugonda received BTech in electronics and communication engineering from JNTU, Hyderabad and MTech in electrical and electronics engineering from JNTUA Anantapur. She is pursuing her PhD in signal processing and communications at JNTUK, Kakinada. She has 15 years of teaching experience and 5 years of research experience. Her research interests include signal processing, image processing and communications.

Corresponding author. Email: Email: [1] rajani.alugonda@jntucek.ac.in

Satya Prasad Kodati has an extensive career spanning 38 years in teaching and 28 years in research. He earned his BTech in electronics and communication engineering from JNTU College of Engineering in 1977, followed by an ME in communication systems from Guindy College of Engineering, Madras University, in 1979,and a PhD from the Indian Institute of Technology, Madras, in 1989. He worked as professor of ECE at JNTUK Kakinada. He authored four textbooks and published over 250 technical papers in national and international conferences and journals. He is a Fellow member of professional bodies like IEEE, IETE, IE (I), and ISTE. His research interests cover a wide range, including communications, signal processing, Image processing, neural networks, and adhoc wireless networks.