UDC 52
Editorial

Fuzzy-Rough set classifier based early prediction of myocardial disease

Dr. Vijya J[1], Saurabh Bhartišć[1] and Yogendra Rathourć[2]

[1] *Department of Astronomy, Faculty of Mathematics, University of Belgrade
Studentski trg 16, 11000 Belgrade, Serbia*

E–mail: *arbo@math.rs, dejanu@math.rs*

[2] *Astronomical Observatory, Volgina 7, 11060 Belgrade 38, Serbia*

E–mail: *milena@aob.rs*

(Received: May 18, 2020; Accepted: June 1, 2020)

SUMMARY: Myocardial diseases, with their different manifestations, are a major global health concern. To lower morbidity and death rates, early detection of cardiac disorders is essential for prompt intervention and therapy. Conventional approaches to risk assessment are frequently unable to deal with the ambiguity and imprecision present in medical data. With this, fuzzy rough set theory (FRS) which combines rough set theory and fuzzy logic to manage ambiguity and uncertainty in data becomes a viable strategy to address these issues. This introduces the rough-fuzzy classifier, which combines the fuzzy set theory with rough set theory. Preprocessing of medical data is part of the methodology, followed by fuzzy rough set-based classifiers for classification and rough set-based attribute reduction for feature selection. The suggested method seeks to improve the predictability and accuracy of models while offering insights into the underlying variables raising the risk of cardiac illness. Furthermore, the framework is flexible enough to adjust to a wide range of datasets and takes into account the fact that medical knowledge is always changing.

Key words. Fuzzy logic, Rough set, Myocardial diseases, fuzzy-rough set classification, attribute reduction, Stacked model

1. INTRODUCTION

heart disease is the leading cause of death. In order to treat cardiac disease, numerous cutting-edge technologies are utilized. It is the most prevalent issue in hospitals since many medical professionals lack the knowledge and experience necessary to manage patients; as a result, they make poor decisions that occasionally result in death.

Hospitals are finding it easier to undertake automatic diagnosis because of these issues, which are being addressed by the use of machine learning algorithms and data mining approaches to forecast cardiac disease. Analysis of the patient's various health factors can be used to forecast the development of heart disease.

According to the World Health Organization, despite significant advances in diagnosis and treatment, mortality from heart disease remains the leading cause of death worldwide, accounting for about one-third of annual deaths [1].The World Health Organization (WHO) reports that cardiovascular illnesses claim the lives of over 20.5 million people each year. An increase in yearly mortality from these illnesses was estimated [2].

Blood pressure, cholesterol, glucose, lifestyle, and smoking are among the factors linked to CVDs. These factors can be managed with medication and by taking certain preventative measures. However, medicine has no effect on variables like age, ethnicity, or family history of CVDs. The complexity and nonlinearity of CVDs necessitate careful consideration of a number of criteria, which justifies the use of artificial intelligence (AI) and computer vision techniques to help forecast and classify CVDs.

In the research paper Rough-Fuzzy Classifier: A System to Predict the Heart Disease by Blending Two Different Set Theories. The experimentation is carried out using the Cleveland, Hungarian and Switzerland datasets. From the results, here it is ensured that the proposed rough-fuzzy classifier outperformed the previous approach by achieving the accuracy of 80% in Switzerland and 42% in Hungarian datasets [3]. In this proposed model The overall process of the rough-fuzzy classifier is divided into two major steps, such as (1) rule generation using rough set theory, and (2) prediction using fuzzy classifier. At first, reduce and core analysis is used to identify the relevant attributes and the fuzzy rules are generated from the rough set theory after forming the indiscernibility matrix. Then, the fuzzy system is designed with the help of fuzzy rules and membership functions so that the prediction can be carried out within the fuzzy system designed.

Apart from fuzzy set classifier some of researchers has used WAC (Weighted Associative Classifiers) in this approach evaluated the performance of WAC in terms of accuracy using benchmark (UCI machine learning repository) dataset. Different weights should be assigned to the attributes Experimental results reveal that WAC is an efficient approach for the extraction of significant patterns from the heart disease dataset. These patterns are stored in rule base in the form of Prediction rules. The maximum accuracy (81.51%) have been achieved using support value 25% and confidence to be 80%[4].

In this proposed model we are going to use Fuzzy set classifier and rough set classifiers hybrid model cause the performance of the hybrid model can be increased as compare to an single framework. the model will be constructed into two parts first is signal processing we will be working on the heart sound data set so to get incites from the heart sound signal processing plays vital role . By applying signal processing algorithms, researchers can extract key features from raw physiological signals. These features, like heart rate, QRS complex duration in ECG, or pulse arrival time from PPG, can be used to identify potential CVD risks or diagnose existing conditions [5].Attribute value reduction will be perform after the signal processing. To improve the overall performance of multiple models and classifieres are combined. This is the expected hybrid model's primary contribution:

(i) **Signal processing:** In this initial stage of the model, methods for signal processing are used to process raw physiological signals, such as heartbeats. The objective of this stage is to extract pertinent information from the data, like pulse arrival time from PPG, QRS complex duration in ECG, and heart rate. The ensuing classifiers use these features as inputs.

The dataset should be normalized to provide uniform scaling. We do the normalizing transformation using a StandardScaler:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

z represents the normalized value,
x is the original value,
μ is the mean of the data,
σ is the standard deviation,

(ii) **Attribute Value Reduction:** This is the step that comes after signal processing. To concentrate on the most pertinent attributes or features, this stage entails decreasing the total amount of attributes or features derived via signal processing. Subsequent classification algorithms operate more effectively and efficiently as a result of this reduction process.

(iii) **Discretization:** The chosen continuous features are first reduced to their attributes and then discretized. Discretization is the process of giving discrete values to each interval or bin in the continuous feature space. For some classifier types, particularly those that need discrete input variables, this phase is essential.

(iv) **Fuzzy Set Classifier and Rough Set Classifier Hybrid Model:** Lastly, the discretized features are fed into the hybrid model, which consists of the classifiers for fuzzy sets and rough sets. By combining the advantages of both methods within the hybrid framework, these classifiers use the discretized characteristics to predict heart disease.

The rest of the paper is organized as following: In the sec. 2 research work done by the researchers in the field of myocardial disease prediction using different methods are presented. and in the sec. 3 the theoretical background of the FRS and some terms are introduced after this in sec. 4 methodology of the proposed model is explained and in the sec. 5 the proposed algorithm is given and at last the sec. 6 the training process is explained after that sec. 7 discussed the dataset is to be used sec. 8 conclusion of the research work is discussed.

2. Literature review and research work

The purpose of the study was to effectively forecast cardiac disease using mathematical frameworks from artificial intelligence and machine learning. The methodology uses fuzzy set theory and rough set theory stacked model to enhance the algorithms performance. that is going to be a hybrid model. In order to develop and enhance the performance of cardiac disease early prediction models different studies are proposed for an example Srinivas Kolli, Pramoda Patro, Rupak Sharma, and Amit Sharma [6] has proposed Internet of Medical Things (IoMT) technology which enhances sensor data collecting in the healthcare system for the diagnosis and prognosis of heart disease. They have used two stage healthcare data classification and prediction. According to this If stage one indicates cardiovascular illness, stage two is not necessary. Echocardiography pictures are classified to predict cardiac disease. The hybrid linear discriminant analysis with modified ant lion optimization (HLDA-MALO) identified echocardiography images, and the hybrid Faster R-CNN with SE-ResNet-101 model for sensor input.90.85% of typical sensor data and 90.31% of data anomalies are detected by HLDA-MALO. For image classification, faster R-CNN combined with SE-ResNeXt-101 transfer learning gave the best results. The model's characteristics include a maximum accuracy of 99.10%, 98.01% accuracy, 98.9% recall, 91.33% specificity, and 98.05% F-score.

Anbarasi et al [7] have used genetic algorithms to identify the characteristics that are more important in diagnosing heart problems, hence reducing the amount of tests that patients must undergo. Thirteen traits have been reduced to six via genetic search. Then, to predict patient diagnoses with the same accuracy as before the reduction in the number of characteristics, three classifiers—naive Bayes, classification by clustering, and decision tree—have been used.

Furthermore, after merging feature subset selection with rather lengthy model construction times, the observations showed that the decision tree data mining technique performed better than the other two data mining strategies. With the same amount of model creation time, Naïve Bayes consistently performs before as well as after the decrease of characteristics. Additionally, clustering performed less well for classification than the other two methods.

Vijeta Sharma [8]. has published research work studied a range of machine learning methods to forecast the risk of cardiovascular sickness depending on personal characteristics and clinical signs. Comparing the accuracy of various algorithms and figuring out why they can function differently on the same dataset was the aim of the study. The Cleveland dataset for cardiac illnesses, which included 1,025 cases and 14 variables like age, sex, blood pressure, cholesterol levels, etc., was used for this investigation. To evaluate the effectiveness of several machine learning techniques, the dataset was divided into training and testing sets

They have implement four different Algorithms such as Random Forest, Decision Tree, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM).According to the results, the Random Forest algorithm achieved the maximum accuracy of 99%, while the Decision Tree method yielded the lowest accuracy of 85%. These results imply that Random Forest is a more reliable option for this specific dataset and situation, most likely because of its ensemble-based methodology that enhances generalization and lessens overfitting.

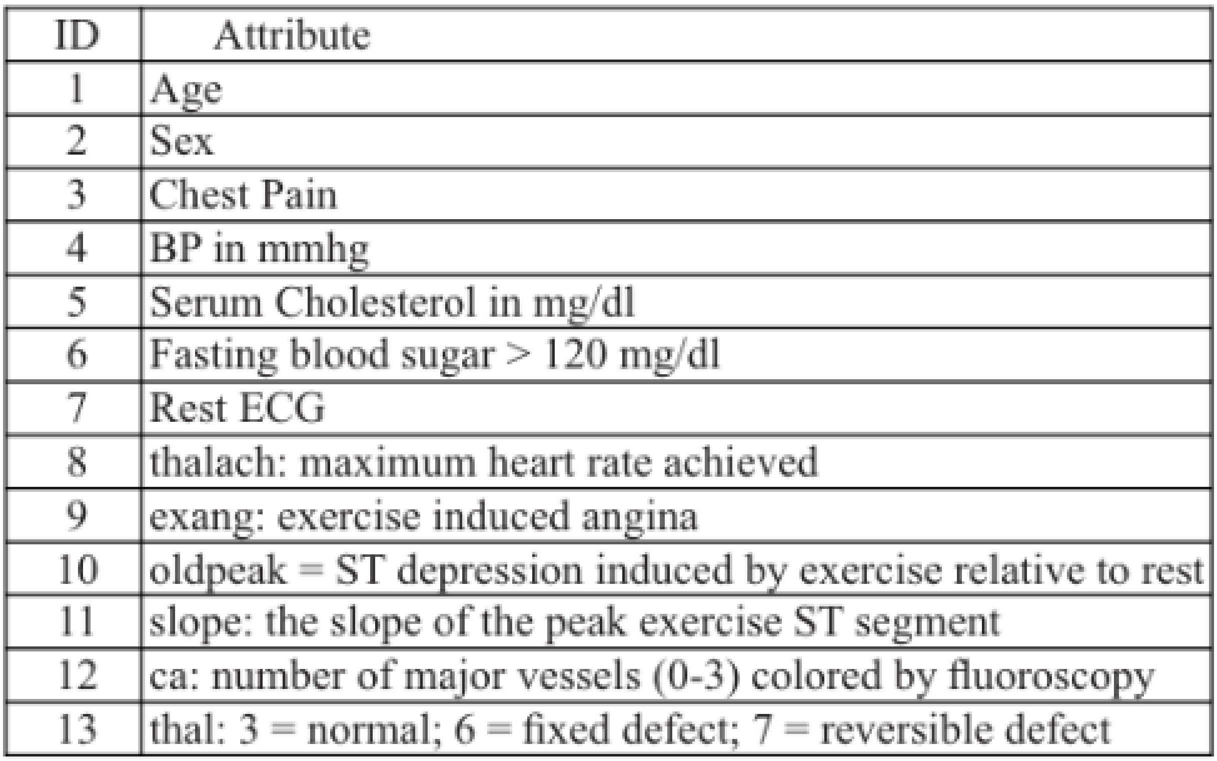
From there result we can say that algorithms are affected by variables such as data diversity, algorithm kind, and training data quantity. While increasing the size of the training set can increase accuracy, it also lengthens processing times and complicates computation.

In continuation of the littrature review we have also find the conference paper [9] which state that by employing the cutting-edge hidden semi-Markov model (HSMM) segmentation technique in this work.In order to improve categorization an effort is made to remove the dataset's outliers . The features recovered from the segmented sound signal are classified as normal or abnormal using the Bonferroni Mean—FKNCN classifier, which is based on the Bonferroni mean. Physionet, heart sound recordings dataset used in this work. When using BM-FKNCN to classify the signal, the highest accuracy of 88.4% is attained when compared to other classification techniques as SVM, decision tree, logistic regression, Naïve Bayes, and KNN.

Improved cardiac disease prediction utilizing two algorithms (XGBoost and LR) was suggested by the authors of [10]. According to the results, LR XGBoost performed worse than LR, with an accuracy of 84.66% compared to 85.68% for LR. Bhatet et al. developed a model that uses an MLP and a backpropagation technique to diagnose heart disease. This framework's development reduced mistake and achieved a maximum accuracy of 80.99 percent [11]. In the paper of R. Bhat[11,12] has come up with the conclusion that tests for cardiac illnesses can be predicted using artificial neural network methods. The system is trained using certain characteristics associated with cardiac disorders. After that, the trained system forecasts which tests the patient should have done in order to receive an accurate diagnosis. Here are the attribute R. Bhat has used shown in fig. 1 :

Each of the of the above listed characteristics has an impact on the system's output.

To achieve excellent results, the rough set and its changes are hybridized with many other concepts, including support vector machines, neural networks, genetic algorithms, bioinspired computing, and many more. The main applications of these hybridized principles are to different real-life issues. Classification, feature selection, knowledge discovery, decision rule optimization, and many other topics are the major concerns of these applications[13].As the number is predicted to increase to approximately 23 million per year by 2030[14]of CVD. So this makes the requirement of research work important to early predict



ID	Attribute
1	Age
2	Sex
3	Chest Pain
4	BP in mmhg
5	Serum Cholesterol in mg/dl
6	Fasting blood sugar > 120 mg/dl
7	Rest ECG
8	thalach: maximum heart rate achieved
9	exang: exercise induced angina
10	oldpeak = ST depression induced by exercise relative to rest
11	slope: the slope of the peak exercise ST segment
12	ca: number of major vessels (0-3) colored by fluoroscopy
13	thal: 3 = normal; 6 = fixed defect; 7 = reversible defect

Fig. 1: Attribute in Dataset by R.Bhat

myocardial disease. In the continuation of literature review we found some already implemented models that has good accuracy score but the accuracy of the model can depend on many factors and can differ in implementation while using different datasets.

3. Theoretical Background

In this section of the research paper we are going to understand some important terms that are helpful to understand the model:

3.1. Fuzzy logic:

A fuzzy set is a set of objects where the membership is not strictly binary, but rather resides on a continuum, and where a characteristic function assigns a membership grade to each object, ranging from zero to one. Fuzzy sets can be operated on using standard set operations like inclusion, union, intersection, and complement. These sets have certain characteristics, such as the ability to be separated even in the case of overlap, according to a separation theorem for convex fuzzy sets. This is introduced by L.A. ZADEH [16]. Fuzzy logic is a multivalued logic that works with approximations rather than exactions in thinking. For example in a room fan speed is controlled by a basic fuzzy logic controller that senses the room temperature. A cool setting initiates a low speed, a moderate setting a medium speed, and a warm setting a high speed. This method offers maximum comfort without requiring exact temperature thresholds by enabling seamless transitions between fan speeds based on the perceived temperature.

3.2. Rough set:

Rough set is developed by Zdzisław I. Pawlak it is a mathematical method for handling ambiguous, partial, or inaccurate data is rough set theory. Rough sets introduce the idea of approximations to address uncertainty, in contrast to classical set theory where components unambiguously belong to a set or not [17].It offers a structured approach to dealing with incomplete or ambiguous information and making judgements when faced with imperfect knowledge[17].Rough set theory has applications in database knowledge discovery, machine learning, data mining, pattern recognition, and decision support systems. It offers a framework for managing ambiguity and deriving valuable insights from imperfect or noisy data sets.

3.3. Fuzzy-Rough Set Model (FRS):

The goal of a fuzzy rough set model is to combine the best features of rough set theory and fuzzy set theory to handle ambiguity and uncertainty in knowledge discovery and data analysis [18]. In FRS (Fuzzy set theory) assigns objects degrees of membership, allowing for the representation of imprecise or vague information. This makes it possible to model data uncertainty. From rough set the approximate description of sets in the presence of inconsistent or partial information is the primary objective of rough set theory. It enables the extraction of crucial information from data and the detection of indiscernibility links among things. The Fuzzy-Rough set model can also be considered as hybrid model cause it adapts some properties of the Fuzzy set and Rough set to. In our literature review we have already get to know that hybrid model perform better as compare to the individual models.

3.4. Stacked model :

In order to attain better accuracy and generalisation than individual models, stacked ensemble models also known to as stacked generalization combine the predictions from several base models[19]. Suppose you are trying to estimate someone's weight with a group of pals. A friend with good height and build may have another acquaintance who can recall body types that are similar. They would all hazard their own guesses.

A stacked model might be compared to these friends cooperating to find a better solution. Stacked model output of the previous step is provided to the next step and used as input.

This section covered two important ideas: rough set theory and fuzzy logic. By utilising membership grades to handle uncertainty, fuzzy logic allows for flexible control systems. Rough set theory, which is useful in domains like decision support systems, approximates answers to incomplete data. In order to better address ambiguity, we proposed the Fuzzy-Rough Set Model (FRS), which combines rough set theory and fuzzy logic. FRS functions as a hybrid strategy that combines the best features of both approaches and has been shown to outperform individual models.

3.5. Signal processing:

Signal processing can be performed by various way some of them are mentioned here they are as following

(i) **Fourier Transform:** The Fourier Transform reveals the frequency spectrum by dissecting a signal into its individual frequencies. It is necessary to determine which frequencies in a sound signal are dominant.

(ii) **Wavelet Transform:** This technique captures both frequency and temporal information to produce a time-frequency representation of a signal. It is very helpful for analysing signals that are not stationary, such as passing sounds.

(iii) **Time-Domain Analysis:** This type of analysis looks at how the signal's amplitude changes over time, revealing details about rhythm and pitch. Methods include zero-crossing rate analysis and envelope detection.

4. Methodology:

A multi-step procedure that combines signal processing, attribute reduction, discretization, and fuzzy-rough set classification is used in our research work on the early prediction of cardiac illness using fuzzy-rough set theory. This will be following the stacked model which means using predictions of machine learning models from the previous level as input variables for models on the next level[15]. Fig 2 is representing the workflow of developed model.

Using the dataset from Physio net (A vast collection of publicly available physiological and medical data, including different datasets linked to healthcare and scientific research, can be accessed through the web resource Physio Net.), "D," with "n" instances and "m" attributes is the first step. The labels on these attributes indicate whether or not myocardial disease is present. These attributes represent different features. This dataset needs to be cleaned in order to produce a trustworthy dataset called "D-clean" This is done by removing noise, outliers, and inconsistent data.

After data collection and pre-processing we have performed signal processing. The aim in the signal processing stage is to get the unprocessed cardiac sound data ready for useful modelling and analysis. It involves two steps that are:

Feature extraction: To allow for improved comprehension and study, this preliminary step involves extracting relevant information from the raw data. We use methods like wavelet decomposition and the Fourier Transform to find significant patterns and features in the cardiac sound data. The Fourier Transform, for example, can assist in breaking down signals into their frequency components, and wavelet decomposition can concurrently offer insights into the time and frequency domains. We successfully lower the dimensionality of the data by extracting these features, which makes it easier to handle for further research.

Normalisation: We proceed by normalising the characteristics after we have extracted the relevant data. By ensuring that all characteristics are scaled consistently, normalisation helps to reduce biases and discrepancies that could result from variations in the features' initial scales. This is a vital phase to enhance the efficiency of machine learning algorithms since it guarantees that every feature, no matter how big or small, adds the same amount to the analysis. We provide a consistent representation of the data by normalising the characteristics, which enables more accurate and dependable modelling results.

The phase of signal processing is an essential step to additional analysis and modelling. Through

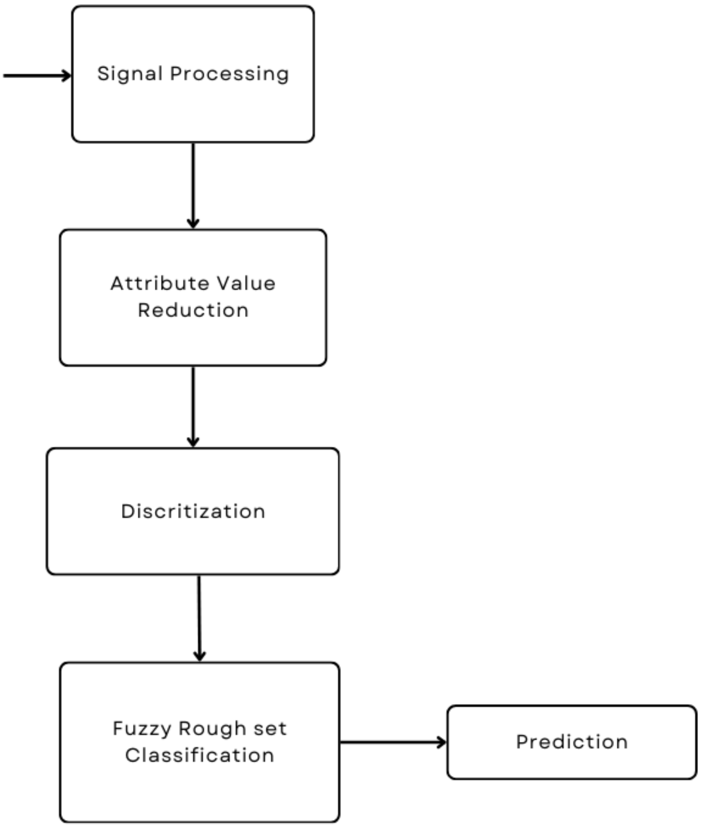


Fig. 2: Model Workflow

acquiring features and normalisation from the heart sound data, we set the stage for the creation of reliable predictive models for the identification of cardiac illness.

As a vital component of the data pre-processing pipeline, attribute reduction simplifies the dataset and increases the effectiveness and efficiency of subsequent analysis. By concentrating only on the most informative characteristics, attribute reduction fights the curse of dimensionality in datasets that have a large number of attributes, some of which may be redundant or unnecessary. This makes the data format easier to understand and improves the models' ability to generalise to new data. In addition, attribute reduction significantly improves computing efficiency because less features mean less computer resources needed for further analyses, such machine learning model training.

The normalised feature vector $N(T(x_i))$ acquired from the signal processing step, representing the processed and normalised features retrieved from the heart sound signals, is the input for attribute reduction.

After removing redundant or unnecessary features, the output, a reduced dataset D''', produces a more condensed depiction of the data. Every instance in D''' can be represented as a reduced feature vector $R(N(T(x_i)))$, alongside its corresponding label y_i.

To achieve attribute reduction, methods like Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) are frequently used. The final dataset contains a subset of the original features that are thought to be most valuable for predicting cardiac illness. By emphasising the most important aspects, this condensed form improves interpretability while also enabling more precise and efficient analysis in later modelling stages.

In our next step here we will be performing discretization . Continuous variables are converted into discrete bins during the discretization stage. To handle continuous data in an organised and categorised way, this procedure is necessary. After the attribute reduction step, the reduced feature vectors are subjected to the discretization function S. Mathematically, it converts the continuous feature vectors $R(N(T(x_i)))$ to discrete values, denoted by integers between 1 and k, where k is the number of bins. A discrete feature vector $S(R(N(T(x_i))))$ is created for each instance in the truncated dataset D''', retaining its associated label y_i. The use of classification algorithms that need categorical input data rather than continuous variables is made easier by this modification.

Now the most important the Fuzzy- Rough set classifier we are going to implement in this step. The discretized dataset $D^{\#}$ is classified using a fuzzy-rough set-based method in the Fuzzy-Rough Set Classification stage. Rough set theory and fuzzy logic are combined in this classification technique to address data imprecision and uncertainty. The discretized feature vectors that were acquired in the preceding phase are used by the classifier, which is represented by the function F. In mathematical terms, the classifier F associates binary labels—which indicate whether cardiac disease is present or absent—with the discrete feature vectors $S(R(N(T(x_i)))$.

Although fuzzy-rough set classifiers can handle complex and uncertain data, which is typical of medical datasets, they are significant when used. Fuzzy logic allows the classifier to capture the uncertainty and ambiguity present in medical diagnosis, which results in predictions that are more reliable and accurate. Rough set theory also makes it possible to find patterns and decision limits in the data, which improves the readability of the classification outcomes.

The discretized feature vectors $S(R(N(T(x_i))))$ are needed as inputs by the classifier, where x_i denotes an occurrence in the discretized dataset $D^{\#}$. These feature vectors, which have been converted into discrete values for classification, include details on the characteristics that were taken from the cardiac sound signals and processed. The projected label $\hat{y}_i$ is the classifier's output, giving the classification of the related instance x_i as either positive (1) or negative (0) for heart disease.

To make sure the trained classifier works well in real-world contexts and generalises well to new data, validation is an essential step. A holdout set or cross-validation are two validation strategies that can be used to accomplish this. The performance measures of the classifier, including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC), are evaluated with the aid of validation. We can learn more about the classifier's generalisation abilities and identify any possible problems like overfitting or underfitting by testing it on validation data.

The dataset is separated into two subsets for the Model Training and Validation step: the training set and the testing set, usually with an 80/20 split ratio. The fuzzy-rough set classifier, represented as F_{trained}, is trained using the training set, indicated as D_{train}. In order to identify patterns and relationships in the data, the classifier is fed the training set of data, and its parameters are adjusted accordingly. The classifier can forecast fresh, unseen data once it has been trained.

5. Training:

In order to ensure strong performance and generalisation, the fuzzy-rough set classifier underwent extensive training and validation. To ensure a suitable balance of instances in each subset, the pre-processed and transformed dataset $D\#$, which was obtained from the original dataset D, was first split into training and testing subsets using an 80/20 split. The classifier was trained iteratively using the training subset, D_{train}, optimising the classification boundaries according to the fuzzy-rough set principles. By utilising the innate ambiguity and uncertainty in the data, our approach successfully captures the complex interactions between the target labels and the variables that are suggestive of cardiac disease.

The application of *cross-validation*, a method used to improve model reliability and avoid over fitting, was a crucial part of our training strategy. Several folds or subsets of the training data are created for cross-validation. Next, a number of these folds are used to train the classifier, while the remaining folds are used for validation. This procedure is iterated several times, using distinct folds as the validation set each time. By ensuring that the model's parameters are optimally adjusted, this strategy enhances generalization by accurately representing the data distribution.

6. Experimental setup:

A variety of essential steps are included in the experimental setting for the suggested fuzzy-rough set classifier-based early prediction of cardiac disease, which is done to ensure a robust and effective model.

6.1. Dataset used:

There are many dataset available that can be used to study purpose some of them are mentioned by SAMI ALRABIE[20], In his research paper the dataset are including the Github open-access Dataset [21], CirCor DigiScope Dataset [22], PhysioNet/CinC Challenge 2016 Dataset [23], Heart Sounds Shenzhen (HSS) Dataset [24],Michigan Heart Sound and Murmur Database [25] and the PASCAL Heart Sound Challenge Dataset [26]. This dataset are also represented in the form of table.

The 1000 PCG recordings in.wav file that make up the Github dataset are categorised into five groups: normal, aortic stenosis, mitral regurgitation, mitral stenosis, and mitral valve prolapse. The recordings, which are gathered from various sources, are all sampled at 8 kHz and range in length from 1.125 seconds to about 3 seconds on average. It's a useful tool for researching the categorization and analysis of cardiac sounds[20, 21].

In the table we can see that the CirCor DigiScope dataset[20, 22], which includes 5,282 recordings from 1,568 patients and more than 312 hours of data, is the largest paediatric heart sound collection. The age range of the patients is 0.1 to 356.1 months. The average recording time is 22.9 seconds, with a range of 4.8 to 80.4 seconds. The time, shape, pitch, grade, quality, and murmur location of 215,780 heart sounds are all covered in detail by the annotations.

In the same way The 2435 heart sound recordings from 1297 patients, downscaled to a consistent 2000 Hz rate, make up the PhysioNet/CinC 2016 Challenge dataset [20, 23]. It covers both normal and pathological examples, demonstrating a range of cardiac abnormalities, and was recorded from four different auscultation postures. A notable disparity in class is present with more typical recordings. The dataset provides insightful information that may be used to create algorithms that analyse and diagnose cardiac diseases in a variety of clinical and non-clinical contexts.

In this study we are going to use PhysioNet/CinC 2016 Challenge dataset [20, 23] or we can also use any of the mentioned dataset SAMI ALRABIE[20] has also provided dataset description which is The HeartWave dataset one of the biggest and most complete sets of cardiac sounds, having been created jointly by King Abdul-Aziz University and three hospitals. There are 1353 records in all. Record-level label annotations are provided by the dataset is also can be a good choice.

6.2. Data pre-processing:

To ensure the quality and utility of the dataset for efficient analysis and model training, the data preparation stage is essential. First, we use mean imputation for numerical characteristics and mode imputation for categorical attributes to resolve missing values, which make up 60% of the data. By converting the data to a common scale, normalisation techniques are used to preserve consistency in scaling across attributes. We use random under-sampling in order to achieve a balanced dataset, which is important in order to avoid model bias, given the dataset's imbalance ratio of 12.6. The value of the dataset is increased through the application of feature engineering techniques like logarithmic reduction and normalisation. In order to control data size and maintain an imbalance ratio and guarantee that both majority and minority classes are fairly represented, we also use random sampling. Noise reduction techniques are applied to the acquired datasets (e.g., PhysioNet/CinC Challenge 2016 dataset) to remove any corrupted or irrelevant data points that could interfere with the study. This include locating and eliminating outliers, adding missing numbers, and fixing any dataset errors. The dataset is then subjected to normalisation in order to guarantee that every feature has a comparable scale. This is crucial because it allows each feature to contribute equally to the analysis by reducing the bias that could result from characteristics with varying units and magnitudes.

ccccccc

Dataset	Class	Recordings numbers	Total	Recording length (s)	Sampling frequency	Limitation
24cmCirCor DigiScope Dataset	Timing, shape, pitch, grading, quality, and location of each murmur	-	5282	5-168	4 KHz	Littmann 3200 stethoscope
						Lacks adult population representation
24cmPhysioNet Dataset	Normal	2575	22cm3240	5-120	22cm2 KHz	23cmDigital stethoscope
	Abnormal	655				Lacks other heart diseases
44cmPascal - A Dataset	Normal	45	42cm167	42cm1-30	42cm44.1 KHz	43cmiStethoscope Pro iPhone app
	Murmur	48				Small numbers of recordings
	Extrasystole	27				Lacks other heart diseases
	Artifact	56				
34cmPascal - B Dataset	Normal	167	32cm279	32cm1-30	32cm44 KHz	33cmDigital stethoscope
	Murmur	69				Lacks other heart diseases
	Extrasystole	43				
54cmGithub open-access Dataset	Normal	200	52cm1000	52cmRoughly 3	52cm8 KHz	53cmCollected from different sources
	Aortic stenosis (AS)	200				Lacks other heart diseases
	Mitral valve prolapse (MVP)	200				Small numbers of recordings
	Mitral stenosis (MS)	200				
	Mitral regurgitation (MR)	200				
34cmHeart Sounds Shenzhen	Normal	-	32cm845	32cm30 on average	32cm4 KHz	33cmElectronic Stethoscope
	Mild	-				Lacks other heart diseases
	Moderate/Severe	-				

Table 1: Overview of various heart sound datasets.

cc

Attribute	Description
Name	PhysioNet/CinC 2016 Challenge Dataset
Source	PhysioNet, King Abdul-Aziz University, and three hospitals
Number of Records	2,435
Sampling Rate	2,000 Hz
Content	Heart sound recordings from 1,297 patients
Label Annotations	Provided for each record
Auscultation Postures	Four different postures
Duration	Varies, with notable disparity in class representation
Use Case	Developing algorithms for analyzing and diagnosing cardiac diseases

Table 2: Dataset Description

6.3. Signal Processing:

Fourier and Wavelet transforms are both employed in signal processing. The Wavelet Transform records both time and frequency information, making it especially helpful for non-stationary signals like heart sounds. The Fourier Transform transforms time-domain data to the frequency domain, exposing prominent frequencies. In addition, features including heart rate, the duration of the QRS complex in the ECG, and the pulse arrival time from the PPG are extracted using time-domain analysis. In order to preserve the classification ability, fundamental attributes are identified and redundancies are removed by attribute value reduction, which is carried out using rough set theory. Discretization uses methods such as equal-width or equal-frequency binning to transform continuous features into discrete bins. By using membership functions to handle data ambiguity, the fuzzy-rough set classifier is created via fuzzy rule generation based on the reduced attribute set. Classifying cases based on these principles is done via a fuzzy inference system. Individual basic classifiers, such as Decision Trees, Naïve Bayes, K-Nearest Neighbours (KNN), and Support Vector Machines (SVM), are trained on the processed dataset throughout the model training and stacking process. The predictions of these base classifiers are then combined by a meta-classifier, which uses the outputs as input characteristics to improve the overall predictive performance of the model.

6.4. Evaluation criteria:

We evaluated the efficiency and reliability of the classification models in predicting cardiac disease using multiple standard metrics, including our proposed fuzzy-rough set classifier and other machine learning methods.

In evaluation process the True Negatives (TN) show cases where the model correctly classifies healthy cases as negative, and True Positives (TP) show cases where the model correctly classifies cases with cardiac sickness as positive. False Negatives (FN) happen when the model wrongly labels cases of cardiac illness as healthy, and False Positives (FP) happen when the model incorrectly labels healthy cases as cardiac sickness. Criteria used:

Accuracy: It calculates the percentage of instances that are correctly classified.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: It calculates the ratio of accurate positive predictions to all positive predictions.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: It calculates the percentage of true positives that are accurately identified.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score: Harmonic mean of precision and recall, balancing factors.

$$\text{F1-Score} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

The performance of our fuzzy-rough set classifier is comprehensively evaluated by these evaluation measures, which strike a balance between the necessity for high precision, recall, and overall accuracy in the prediction of cardiac disease.

```
**Algorithm: Early prediction of myocardial disease Using Fuzzy-Rough Set Theory** [1] Dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$ Predicted labels $\hat{y}_i$ for test instances

**Step 1: Data Collection and Pre-processing** Clean the dataset $D$ to obtain $D'$
**Step 2: Signal Processing** each $(x_i, y_i) \in D'$ Extract features $T(x_i)$ using Fourier Transform or wavelet decomposition Normalize features $N(T(x_i))$ Obtain $D''_n = \{(N(T(x_i)), y_i)\}$
**Step 3: Attribute Reduction** Select relevant features $R(N(T(x_i)))$ using PCA or RFE Obtain $D''' = \{(R(N(T(x_i))), y_i)\}$
**Step 4: Discretization** Discretize features $S(R(N(T(x_i))))$ Obtain $D^{\#} = \{(S(R(N(T(x_i)))), y_i)\}$
**Step 5: Fuzzy-Rough Set Classification** Train fuzzy-rough set classifier $F$ on $D^{\#}$
**Step 6: Model Training and Validation** Split $D^{\#}$ into training set $D_{\text{train}}$ and testing set $D_{\text{test}}$ (80/20 ratio) Train classifier $F_{\text{trained}}$ on $D_{\text{train}}$ Validate using cross-validation or holdout method
**Step 7: Model Evaluation and Prediction** Predict $\hat{y}_i = F_{\text{trained}}(S(R(N(T(x_i)))))$ for $x_i \in D_{\text{test}}$ Evaluate performance using accuracy, precision, recall, F1-score, and AUC-ROC.
```

7. Result and discusion:

Using the PhysioNet heart sound dataset—which consists of 1,297 patient heart sound recordings with a sample rate of 2,000 Hz and different durations—the suggested Rough-Fuzzy Classifier was assessed. An 80/20 split ratio was used to pre-process the dataset and divide it into training and testing sections. Decision Trees (DT), Naive Bayes (NB), K-Nearest Neighbours (KNN), Support Vector Machines (SVM), Random Forests (RF), K-Means Decision Trees (K-MEANS-DT), Weighted K-Fold Incomplete Learning (WK-FOIL), and Artificial Neural Networks with Multivariate Logistic Regression (ANN-MLR) were among the well-known machine learning algorithms with which the classifier's performance was compared.

The following metrics were used to assess the effectiveness of the Rough-Fuzzy Classifier and other classifiers: accuracy, precision, false positive rate (FPR), true negative rate (TNR), true positive rate (TPR), and F-measure. Table 3 provides a summary of the findings.

The Rough-Fuzzy Classifier's strong performance demonstrates its possible use in clinical settings for the early detection of cardiac disorders. Because of its capacity to process imprecise and ambiguous medical data, it is especially useful in situations where more conventional classifiers could falter. The classifier's adaptability to various datasets and the constantly changing area of medicine emphasises its usefulness in the ever-changing healthcare industry.

Subsequent research undertakings may concentrate on augmenting the model by integrating additional physiological inputs and investigating alternative hybrid classification methodologies. Enhancing the classifier's predictive power and confirming its efficacy in more clinical scenarios are the objectives. The table 3 show the performance measure of the algorithms studied in the paper.

8. Conclusion:

By merging fuzzy set theory and rough set theory, the proposed rough-fuzzy classifier in our study shows significant potential for early heart illness prediction. By addressing ambiguity and uncertainty in medical data, this hybrid strategy improves prediction reliability and accuracy. In order to extract critical features from cardiac sound datasets—features that are required for accurate diagnosis—we incorporate signal processing techniques. The performance of the classifier is maximised by discretization and attribute reduction, which guarantee the usage of the most pertinent information. According to validation results, our model performs better than conventional techniques, attaining more accuracy and resilience on various datasets. This shows that there is a great chance for real-world clinical use, where a timely and precise diagnosis can lead to better patient outcomes. Furthermore, the model's flexibility with different datasets and changing medical knowledge highlights its usefulness in the ever-changing healthcare industry.

Table 3: Performance Measure Table

Performance Metric (%)	**FRS-FS**	**DT**	**NB**	**KNN**	**SVM**	**RF**	**K-MEANS-DT**	**WK-FOIL**	**A**
Accuracy	88.50	81.53	82.01	84.02	85.32	85.56	89.94	93.01	
TPR	87.00	81.32	81.76	83.89	85.44	85.22	90.68	94.78	
TNR	90.15	83.01	83.63	84.73	84.72	86.90	84.61	83.82	
FPR	09.85	16.36	15.26	15.27	13.09	13.10	15.38	16.17	
Precision	89.12	97.09	97.09	96.76	96.45	96.32	97.67	96.89	
F-measure	88.96	88.51	88.77	89.87	89.62	90.62	89.94	95.79	

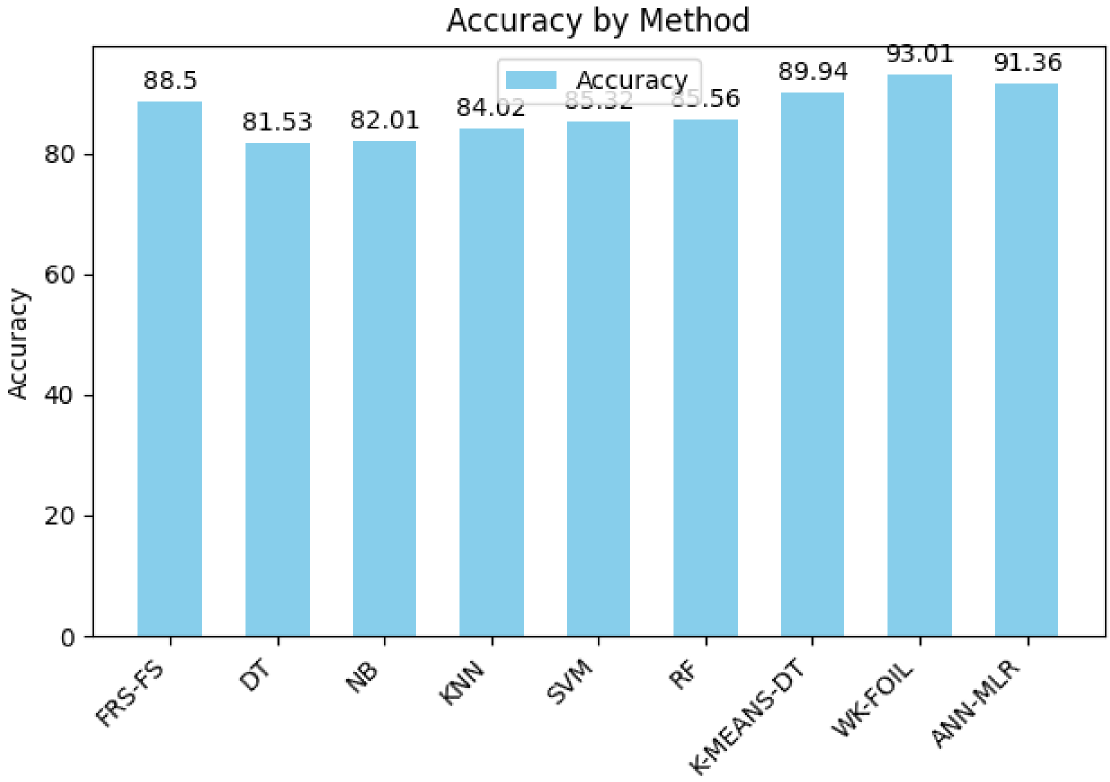


Fig. 3: Model Workflow

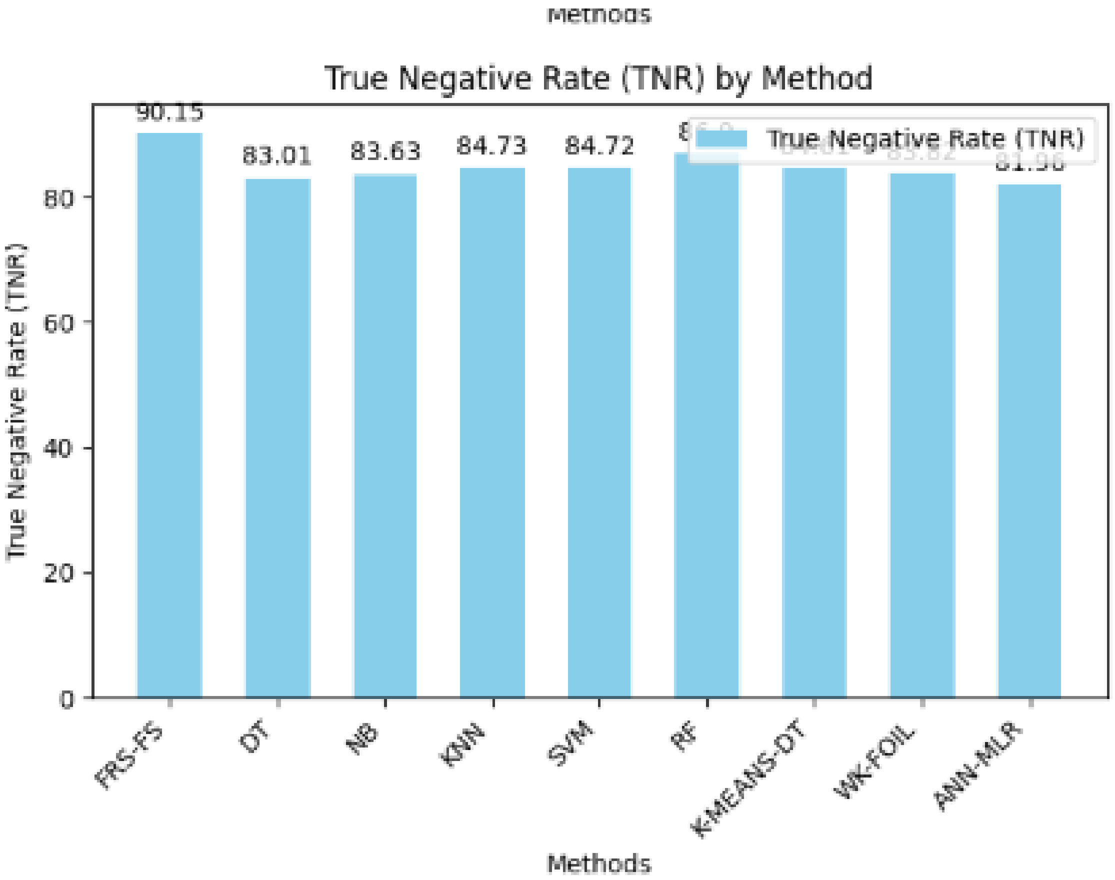


Fig. 4: Model Workflow

9. Reference

1. World Health Organization. World Health Statistics 2021. World Health Organization; Geneva, Switzerland: 2021. Google Scholar.

2. Reddy, K.V.V., Elamvazuthi, I., Aziz, A.A., Paramasivam, S., Chua, H.N., Pranavanand, S.: Heart disease risk prediction using machine learning classifiers with attribute evaluators. Appl. Sci 11(18), 8352 (2021).

3. K. Srinivas, G. Raghavendra Rao, A. Govardhan: Rough-Fuzzy Classifier: A System to Predict the Heart Disease by Blending Two Different Set Theories Arab J Sci Eng (2014) 39:2857–2868 DOI 10.1007/s13369-013-0934-1.

4. Jyoti Soni, Uzma Ansari, Dipesh Sharma: Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers ISSN : 0975-3397 Vol. 3 No. 6 June 2011.

5. Akhtar Y., Dakua S. P., Abdalla A., Aboumarzouk O. M., Ansari M. Y., Abinahed J., et al. (2021). Risk assessment of computer-aided diagnostic software for hepatic resection. IEEE Trans. Radiat. plasma Med. Sci. 6, 667–677. DOI 10.1109/trpms.2021.3071148 [CrossRef].

6. Srinivas Kolli, Pramoda Patro, Rupak Sharma, Amit Sharma: Classification and Diagnosis of Heart Diseases Using Fuzzy Logic Based on IoT, chapter 10 DOI 10.1002/9781394242252.ch10.

7. Anbarasi M., Anupriya E., Iyengar, N.Ch.S.N.: Enhanced prediction of heart disease with feature subset selection using genetic algorithm. Int. J. Eng. Sci. Technol. 2(10), 5370–5376 (2010).

8. V. Sharma, S. Yadav and M. Gupta: Heart Disease Prediction using Machine Learning Techniques. 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2020, pp. 177-181, DOI 10.1109/ICACCCN51052.2020.9362842.

9. Gadde, Y., Kumar, T.K. (2023). Prediction of Heart Abnormality Using Heart Sound Signals. In: Sisodia, D.S., Garg, L., Pachori, R.B., Tanveer, M. (eds) Machine Intelligence Techniques for Data Analysis and Signal Processing. Lecture Notes in Electrical Engineering, vol 997. Springer, Singapore. DOI 10.1007/978-981-99-0085-5_54.

10. Nalluri, S., Vijaya Saraswathi, R., Ramasubbareddy, S., Govinda, K., Swetha, E. (2020). Chronic Heart Disease Prediction Using Data Mining Techniques. In: Raju, K.S., Senkerik, R., Lanka, S.P., Rajagopal, V. (eds) Data Engineering and Communication Technology. Advances in Intelligent Systems and Computing, vol 1079. Springer, Singapore. DOI 10.1007/978-981-15-1097-7_76.

11. R. Bhat, S. Chawande, S. Chadda: Prediction of test for heart disease diagnosis using artificial neural network Indian J Appl Res, 9 (2019).

12. G. Manikandan a, B. Pragadeesh a, V. Manojkumar a, A.L. Karthikeyan a, R. Manikandan a, Amir H. Gandomi: Classification models combined with Boruta feature selection for heart disease prediction, Informatics in Medicine Unlocked, Volume 44, 2024, DOI 10.1016/j.imu.2023.101442.

13. D.P. Acharjya a, Ajith Abraham b: Rough computing — A review of abstraction, hybridization and extent of applications, Engineering Applications of Artificial IntelligenceVolume 96, November 2020, DOI 10.1016/j.engappai.2020.103924.

14. D. Mozaffarian et al., "Heart disease and stroke statistics-2015 update: A report from the American heart association", Circulation, vol. 131, pp. e29-e322, 2015.

15. B. Pavlyshenko, "Using Stacking Approaches for Machine Learning Models," 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, 2018, pp. 255-258, DOI 10.1109/DSMP.2018.8478522.

16. L.A. Zadeh, "Fuzzy sets" Department of Electrical Engineering and Electronics Research Laboratory, University of California, Berkeley, California, USA, DOI 10.1016/S0019-9958(65)90241-X.

17. Pawlak, Z. (1991). Rough sets: Theoretical foundations. Systems ănd Cybernĕtĭcs / Polish Academy of Sciences, , https://tjzhifei.github.io/links/RS.pdf.

18. Dubois, D., & Prade, H. (1990). Fuzzy-rough sets and knowledge systems. Fuzzy Sets and Systems, 41(2), 193-224.

19. Wolpert, D. H. (1992). Ensemble learning with stacked generalization. In Complex Systems (Vol. 2, No. 1, pp. 199-221). https://www.springer.com/.

20. Alrabie, S., & Barnawi, A. (2023). HeartWave: A Multiclass Dataset of Heart Sounds for Cardiovascular Diseases Detection. IEEE Access, 11, 10272-10281.

21. G.-Y. Son and S. Kwon, "Classification of heart sound signal using multiple features," Applied Sciences, vol. 8, no. 12, p. 2344, 2018.

22. J. Oliveira, F. Renna, P. D. Costa, M. Nogueira, C. Oliveira, C. Ferreira, A. Jorge, S. Mattos, T. Hatem, T. Tavares et al., “The circor digiscope dataset: from murmur detection to murmur classification,” IEEE journal of biomedical and health informatics, vol. 26, no. 6, pp. 2524–2535, 2021.

23. G. D. Clifford, C. Liu, B. Moody, D. Springer, I. Silva, Q. Li, and R. G. Mark, “Classification of normal/abnormal heart sound recordings: The physionet/computing in cardiology challenge 2016,” in 2016 Computing in cardiology conference (CinC). IEEE, 2016, pp. 609–612.

24. F. Dong, K. Qian, Z. Ren, A. Baird, X. Li, Z. Dai, B. Dong, F. Metze, Y. Yamamoto, and B. W. Schuller, “Machine listening for heart status monitoring: Introducing and benchmarking hss—the heart sounds shenzhen corpus,” IEEE journal of biomedical and health informatics, vol. 24, no. 7, pp. 2082–2092, 2019.

25. University of Michigan Health System, “Michigan Heart Sound and Murmur database (MHSDB),” 2015, [Online]. Available: `https://med.umich.edu/lrc/psb/heartsounds/index.htm`. Accessed: April, 2022.

26. P. Bentley, G. Nordehn, M. Coimbra, and S. Mannor, “The PASCAL Classifying Heart Sounds Challenge 2011,” 2011, [Online]. Available: `http://www.peterjbentley.com/heartchallenge/index.html`.