



# **Report - IMA205 Challenge 2023**

**ROCHA PESTANA Thalís**  
Nickname : Giuseppe Camolli

May 08, 2023

**Professors** : Loic Le Folgoc / Pietro Gori

**Institution** : Télécom Paris

## SUMMARY

<b>1. Introduction</b>	<b>3</b>
<b>2. Methodology and Development</b>	<b>4</b>
2.1 Feature Extraction :	4
2.2 Preprocessing :	5
2.2.1 Splitting and Scaling :	5
2.2.2 Verifying Class Imbalance :	6
2.2.3 Feature Importance	6
2.2.4 Verifying Linear Separability and Linear Relationship between Features and Target :	7
2.2.4.1 Method I : Model Performance :	7
2.2.4.2 Method II : Correlation Coefficients :	8
2.3 Models :	8
2.3.1 Random Forest :	9
2.3.2 Stacking :	9
<b>3. Results :</b>	<b>10</b>
4. Conclusion :	13
<b>References</b>	<b>14</b>

# 1. Introduction

Cardiac function analysis is crucial in clinical cardiology for early disease diagnosis, patient management, and therapy decisions. In this report, we present an approach for classifying cardiac magnetic resonance imaging (CMRI) into five diagnostic classes: healthy controls, myocardial infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy, and abnormal right ventricle, with the goal of improving early disease detection and prevention of complications such as heart failure and sudden cardiac arrest.

In order to achieve this goal, we applied a Random Forest Classifier with hyperparameter tuning (Grid Search CV) based on the works developed by Isensee et al. [1], Khened et al.[2] and Wolterink et al. [3], obtaining an accuracy score of 88.571% (private score). To enhance this result, we applied Stacking, an ensemble learning technique that combines multiple base classifiers to improve predictive performance, obtaining an accuracy score of 91.428% (private score). In addition, a voting technique was applied to combine the results of different features selection algorithms to create a more robust and potentially more accurate final set of features, also avoiding overfitting.

The results show a model performance very similar to the performance achieved by the studies provided as reference, indicating that the combined approach of using a Stacking and a voting technique for feature selection has the potential to effectively classify CMRI images.

## 2. Methodology and Development

A dataset of 150 subjects with their MRI images and their corresponding partial segmentations and metadata (subject height and weight) was provided. Data has already been randomly split into a training-validation set (100 subjects) and a test set (50 subjects). Each subject contains 4 .nii files corresponding to the MRI images at end diastole (dilation) and end systole (contraction), pre-segmented (only right ventricle and myocardium) and non-segmented. Each image is a 3D volume containing the heart and adjacent structures [4].

Aiming to focus on the classification task, we used the pre-segmented images filling the left ventricular region. After that, each structure: Right Ventricular, Left Ventricular and Myocardial, duly segmented, were stored in different arrays so that segmented and non-segmented images at the end of diastole and systole were stored in different arrays, in order to make the interpretation of data and functions easier.

### 2.1 Feature Extraction :

The feature extraction process in this study is designed to gather relevant information from the cardiac magnetic resonance imaging (CMRI) data to aid in the classification of various cardiac conditions. The process involves defining several functions to extract specific features from the images, such as left ventricle (LV), right ventricle (RV), and myocardium (M) segmentations, and calculating various morphological and functional parameters.

First, segmentation functions (LV\_segmentation, M\_segmentation, and RV\_segmentation) are implemented to isolate the LV, RV, and M regions from the input images. These functions are applied to each slice of the input image to generate separate segmented images for the LV, RV, and M regions.

Next, a series of functions are defined to calculate various features from the segmented images, such as the volume (compute\_volume), area (compute\_area) and normalized volume (normalized\_volume). Other morphological features are also computed, including mean circularity (mean\_circularity), maximum and minimum circumference (max\_circumference and min\_circumference), and mean and maximum thickness (mean\_thickness and max\_thickness). Functional parameters like ejection fraction (ejection\_fraction) and body mass index (IMC) are calculated using the extracted features. These features provide essential information about the cardiac structure and function, which can be used to differentiate between healthy and diseased conditions. In total 37 features were extracted mainly following the approach of Isensee et al. [1].

Ultimately, a pandas DataFrame consisting of 37 features for each patient was constructed. This DataFrame incorporates the patient ID, volumes, ejection fractions, circumferences, circularity, thickness, and other derived features, such as the ratio of RV to LV volumes and myocardium to LV volumes. Table I displays the features extracted at end-systole and end-diastole for each segmented structure.

Table I - Features calculated at end systole and end diastole.

	<b>LV</b>	<b>M</b>	<b>RV</b>
<b>Volume</b>	x	x	x
<b>Ejection Fraction</b>	x		x
<b>Mean Thickness</b>		x	
<b>Max Thickness</b>		x	
<b>Mean Circularity</b>	x	x	x
<b>Max Circumference</b>	x	x	x
<b>Min Circumference</b>	x	x	x

Source: Produced by the author.

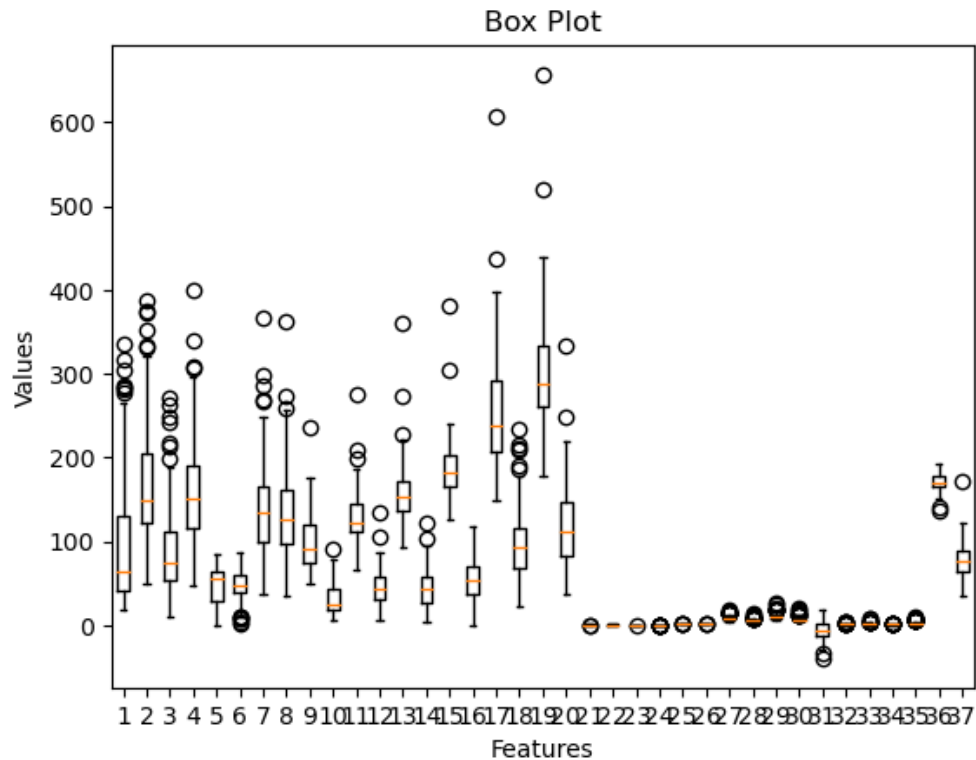
## 2.2 Preprocessing :

### 2.2.1 Splitting and Scaling :

The pandas DataFrame was divided into training and test datasets using the `train_test_split` function, adhering to the proportions prescribed by the Challenge ( $\frac{2}{3}$  for the training set and  $\frac{1}{3}$  for the testing set).

RobustScaler was employed to address the presence of a considerable number of outliers in the data, as evidenced by Figure I. Outliers can significantly impact the performance of machine learning models, and by applying RobustScaler, the model's stability and resistance to such influence can be enhanced. This preprocessing step assists in ensuring that the model is well-prepared to discern underlying patterns within the data and generate accurate predictions for new, unseen data.

Figure I - Boxplot for each feature.



Source: Produced by the author.

### 2.2.2 Verifying Class Imbalance :

An examination for class imbalance was conducted to ascertain that the models possess an equal opportunity to discern patterns from every class. The existence of class imbalance can influence the selection of evaluation metrics and potentially result in suboptimal performance of machine learning models. Nevertheless, upon a straightforward verification, it was observed that the training-validation set exhibited perfect balance, containing 20 samples per class.

### 2.2.3 Feature Importance

In this study, we performed feature importance analysis to identify the most significant features that contribute to the accurate classification of cardiac conditions. By incorporating these essential features into our machine learning models, we aim to build a robust and reliable system capable of providing valuable insights into the diagnosis and treatment of various cardiac pathologies.

Although algorithms like Random Forest can handle high-dimensional data and implicitly perform feature selection during the training process, it is still advantageous to conduct an explicit feature importance analysis for several reasons. Understanding the importance of each feature contributes to the interpretability of the model. By identifying the most important features, we can better comprehend the factors that contribute to the classification of various cardiac conditions.

In addition, feature selection can improve the model's performance by removing irrelevant or redundant features. In this way we can enhance the model's ability to predict outcomes by reducing the noise in the data and concentrating on the most informative features. This simplification can lead to better generalization and higher accuracy in predicting cardiac pathologies.

An extensive feature selection methodology was implemented to determine the most relevant features for our analysis. This comprehensive approach combined several feature selection techniques, including K-Best, Random Forest feature importance, correlation-based feature selection, and mutual information-based feature selection. The selected features from each method were combined, and a voting mechanism was employed to finalize the best features. In this mechanism, the occurrence of each feature across all methods was counted, and only the features that were selected a minimum of three times were retained for further analysis.

#### 2.2.4 Verifying Linear Separability and Linear Relationship between Features and Target :

An alternative approach explored in this study was Principal Component Analysis (PCA). PCA is a widely used statistical technique for dimensionality reduction and data visualization. It operates by transforming the original data into a new coordinate system defined by orthogonal axes, known as principal components. These principal components are linear combinations of the original features, ranked according to the amount of variance they explain in the data. By selecting the top principal components that account for the most variance, it is possible to effectively reduce the dataset's dimensionality while preserving the majority of information.

However, PCA assumes that the principal components are linear combinations of the original features. If the relationships between features and the target variable are highly nonlinear, PCA may not capture the most relevant information. To investigate the suboptimal performance of PCA in this context, the linearity between features and targets was examined using two methods: Model Performance and Correlation Coefficients. This analysis enables a better understanding of the dataset's behavior and informs the choice of appropriate models (linear or nonlinear) for this specific context.

##### 2.2.4.1 Method I : Model Performance :

To assess the linearity between features and the target variable, two types of models were fitted to the data: a linear model (e.g., linear regression) and a nonlinear model (e.g., random forest). If the nonlinear model significantly outperforms the linear model, this suggests that the relationships between features and the target variable

are highly nonlinear. Conversely, if the linear model performs comparably to the nonlinear model, it indicates that the relationships are predominantly linear, and linear models may be sufficient to capture the underlying patterns in the data. In our case the linear model achieved an average mean squared error of 46.49 and the nonlinear model 7.17, suggesting highly nonlinearity.

#### 2.2.4.2 Method II : Correlation Coefficients :

Correlation coefficients are a measure of the strength and direction of a linear relationship between two variables. They can be used to assess the degree of association between features and the target variable in a dataset. The most common correlation coefficient is the Pearson correlation coefficient, which ranges from -1 to 1. A value of 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship between the variables.

In this study, three distinct types of correlation coefficients were calculated for each feature present in the dataset: Pearson, Spearman, and Kendall. Spearman and Kendall correlation coefficients are rank-based measures that exhibit lower sensitivity to outliers and possess the ability to capture monotonic (nonlinear) relationships. Subsequent to computing the correlation coefficients for every feature, the mean and standard deviation were determined for each type of correlation coefficient. The results can be observed in the table below:

Table II - Means and Standard Deviations of three correlation coefficients.

	<b>Pearson</b>	<b>Spearman</b>	<b>Kendall</b>
<b>Mean</b>	0.044	0.0036	0.0052
<b>Standard Deviation</b>	0.189	0.177	0.117

Source: Produced by the author.

## 2.3 Models :

The application of the random forest classifier in the model is justified based on its successful implementation in previous studies focused on automatic cardiac diagnosis, as reported in the paper by Bernard et al. titled "Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?" [5]. Several participants in the Automatic Cardiac Diagnosis Challenge (ACDC) employed random forests for classification tasks after extracting features from their segmentation maps.

For instance, Isensee et al. [1] utilized an ensemble of 50 multilayer perceptrons and a random forest for classification after extracting features from segmentation maps.



Similarly, Khened et al. [2] trained a 100-trees random forest classifier using 11 features, including nine derived from their segmentation map and the patient's weight and height. Wolterink et al. [3] also implemented a five-class random forest classifier with 1,000 decision trees, using 14 features (12 from the segmentation maps and patient weight and height).

These examples demonstrate that random forest classifiers have been successfully employed for cardiac diagnosis tasks in various settings, indicating their potential effectiveness for the model. Additionally, random forests offer several advantages, such as robustness to overfitting, handling of missing data, and the ability to model complex interactions between features. Given these factors, using a random forest classifier in the model is a well-supported choice for automatic cardiac diagnosis.

### 2.3.1 Random Forest :

In the context of this study, the methodology for applying a random forest classifier comprises all the stages cited previously, including data preprocessing, feature extraction, model training, hyperparameter tuning, and model evaluation. The rationale for employing a random forest classifier in this study stems from its various advantages. Random forests help mitigate overfitting by averaging the predictions of multiple decision trees and can handle missing data effectively, either by considering the average value of a feature in the corresponding node or by leveraging the information from other trees in the ensemble.

Another advantage of using a random forest classifier is its ability to capture complex interactions between features. This ability allows the random forest to model complex relationships that may not be captured by simpler models. Furthermore, random forests provide a measure of feature importance, which can be valuable in understanding the most influential factors in the classification task.

To optimize the performance of the random forest classifier, a grid search cross-validation (CV) was employed.

### 2.3.2 Stacking :

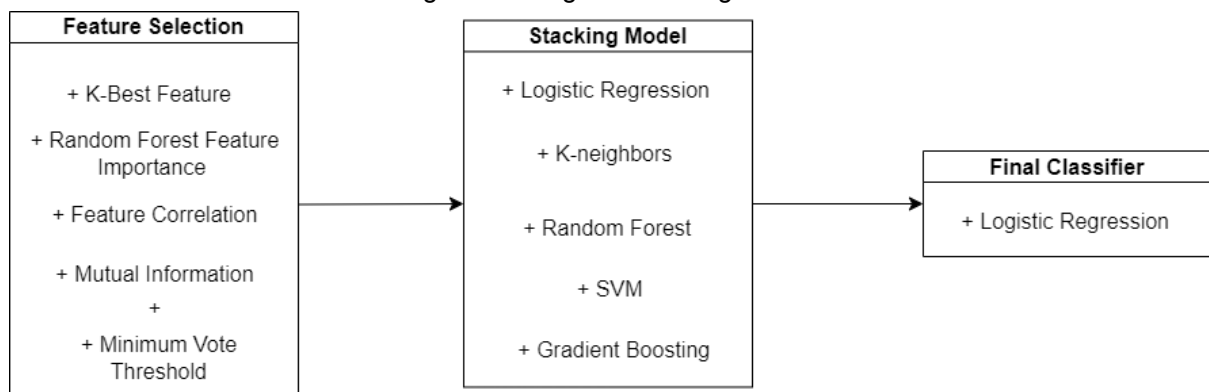
Stacking was applied as an additional ensemble learning technique to further enhance the performance and robustness of the model. Stacking, also known as stacked generalization, involves combining the predictions of multiple base models (also known as level-0 models) using a meta-model (also known as a level-1 model). The goal of stacking is to leverage the strengths of diverse base models, allowing the ensemble to achieve better predictive performance than any single base model.

Each base model may have unique strengths and weaknesses, and combining them through stacking can help overcome their individual limitations. For example, a base

model may perform well on a specific subset of the data or capture specific relationships between features, while another base model may excel in different areas. By combining these base models, stacking can create a more comprehensive and accurate model that benefits from their individual strengths.

Stacking can also reduce the risk of overfitting, as the predictions of multiple base models are combined using a meta-model, which is trained to make a final decision based on the outputs of the base models. Analogously, a grid search CV was applied in each model. The Figure I below illustrates the architecture implemented.

Figure II - Diagram Stacking Model



Source: Produced by the author.

### 3. Results :

All the results were obtained by fixing the random state (equals to 42) of all the models in order to properly compare the results and find the optimal parameters for this specific problem. Concerning the Random Forest classifier, the results presented below show its performance with different features sets and preprocessing methods, using the mean and standard deviation score of K-fold cross validation (Fitting 5 folds for each of 648 candidates, totalling 3240 fits) on the training set and the public/private scores on the Kaggle Challenge:

#### 1. Result with only the features from Khened et al [2] without scaling:

Best parameters found by Grid Search CV:

{'max\_depth': None, 'min\_samples\_leaf': 10, 'min\_samples\_split': 2, 'n\_estimators': 1000}

Cross-validated scores: [0.85 0.95 0.95 0.75 0.85]

Average CV score: 0.8699999999999999

Public Score: 0.8

Private Score: 0.85714

#### 2. Result with Feature Selection (19 features) without Robust Scaling:

Best parameters found by Grid Search CV:

{'max\_depth': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 500}

Cross-validated scores: [1. 0.95 0.95 0.9 0.95]

Average CV score: 0.95

Public Score: 0.86666

Private Score: 0.85714

#### 3. Result with Feature Selection (16 features) and Robust Scaling:

Best parameters found by Grid Search CV:

{'max\_depth': 10, 'min\_samples\_leaf': 5, 'min\_samples\_split': 10, 'n\_estimators': 500}

Cross-validated scores: [1. 0.9 0.95 0.9 0.95]

Average CV score: 0.9399999999999998

Public Score: 0.86666

Private Score: 0.88571

#### 4. Result with only the features from Khened et al[2] with Robust Scaling:

Best parameters found by Grid Search CV:

{'max\_depth': None, 'min\_samples\_leaf': 10, 'min\_samples\_split': 4,  
'n\_estimators': 500}  
Cross-validated scores: [0.75 0.95 0.9 0.75 0.85]  
Average CV score: 0.8400000000000001  
Public Score: 0.8  
Private Score: 0.85714

Due to the substantial computational time required by the stacking model, only the best result, achieved with the following hyperparameters, is presented here:

1. Result with only the features from Khened et al[2] without scaling:

LogisticRegression - Best parameters: {'C': 0.1, 'solver': 'sag', 'tol': 0.01}  
LogisticRegression - Best score: 94.00%  
KNeighbors - Best parameters: {'metric': 'manhattan', 'n\_neighbors': 4,  
'weights': 'distance'}  
KNeighbors - Best score: 90.00%  
RandomForest - Best parameters: {'max\_depth': None, 'min\_samples\_leaf': 1,  
'min\_samples\_split': 3, 'n\_estimators': 2000}  
RandomForest - Best score: 87.00%  
SVM - Best parameters: {'C': 10, 'gamma': 'scale', 'kernel': 'rbf'}  
SVM - Best score: 94.00%  
GradientBoosting - Best parameters: {'learning\_rate': 0.1, 'max\_depth': 3,  
'n\_estimators': 100, 'subsample': 0.5}  
GradientBoosting - Best score: 88.00%  
StackingClassifier - k-fold cross-validation mean score: 93.00%  
StackingClassifier - k-fold cross-validation standard deviation: 7.48%  
Public Score: 0.8  
Private Score: 0.88571




2. Result with Feature Selection (16 features) and Robust Scaling:

LogisticRegression - Best parameters: {'C': 10.0, 'solver': 'saga', 'tol': 0.01}  
LogisticRegression - Best score: 95.00%  
KNeighbors - Best parameters: {'metric': 'manhattan', 'n\_neighbors': 5,  
'weights': 'distance'}  
KNeighbors - Best score: 93.00%  
RandomForest - Best parameters: {'max\_depth': None, 'min\_samples\_leaf': 1,  
'min\_samples\_split': 2, 'n\_estimators': 1000}  
RandomForest - Best score: 94.00%  
SVM - Best parameters: {'C': 1, 'gamma': 'auto', 'kernel': 'rbf'}  
SVM - Best score: 92.00%  
GradientBoosting - Best parameters: {'learning\_rate': 0.2, 'max\_depth': 5,  
'n\_estimators': 50, 'subsample': 0.5}

GradientBoosting - Best score: 94.00%  
StackingClassifier - k-fold cross-validation mean score: 91.00%  
StackingClassifier - k-fold cross-validation standard deviation: 5.83%  
Public Score: 0.8  
Private Score: 0.91428

The consistency of this result was demonstrated by its attainment in three consecutive instances on the private leaderboard of the Kaggle challenge, as can be observed in the figure below:

Figure II - Best results for the Kaggle Challenge.

	<b>results2350.csv</b> Complete (after deadline) · 3d ago	<b>0.91428</b>	<b>0.8</b>
	<b>results2317.csv</b> Complete (after deadline) · 3d ago	<b>0.91428</b>	<b>0.8</b>
	<b>results2315.csv</b> Complete (after deadline) · 3d ago	<b>0.91428</b>	<b>0.8</b>

Source: IMA205 Challenge 2023 [4].

## 4. Conclusion :

In the context of the Random Forest classifier, an interesting observation can be made regarding the mean CV, public, and private scores. It is evident that when Feature Selection is implemented, the average CV score increases, yet the public and private scores decrease. This suggests that implementing Feature Selection may lead to overfitting. Potential reasons for this could be feature selection bias and/or the inclusion of noisy or unreliable features in the selected subset. In either case, if the chosen features contain noise or are not genuinely relevant for the prediction task, the model may learn to depend on this noise or spurious correlations to make predictions. Nevertheless, it is crucial to emphasize that the model with 16 selected features and scaling still presented the best accuracy on the private subset.

The StackingClassifier demonstrates its effectiveness as an ensemble method. In both scenarios, the model achieved competitive performance compared to the individual classifier (Random Forest), showcasing its capability to efficiently combine the strengths of different classifiers to enhance overall performance. However, some challenges, such as high computational cost and lengthy processing time (approximately 50 minutes), should be noted.

It is important to highlight that due to the unavailability of private scores, the best public score was chosen at the time of the challenge's conclusion. However, after the release of private scores, the model (among the various models tested) that generated this specific score was found to be overfitted. The best score has been provided above.

## References

- [1] Isensee, F., Jaeger, P.F., Full, P.M., Wolf, I., Engelhardt, S. and Maier-Hein, K.H., 2018. Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features. In Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers 8 (pp. 120-129). Springer International Publishing.
- [2] Khened, M., Alex, V. and Krishnamurthi, G., 2018. Densely connected fully convolutional network for short-axis cardiac cine MR image segmentation and heart diagnosis using random forest. In Statistical Atlases and Computational Models of

the Heart. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers 8 (pp. 140-151). Springer International Publishing.

[3] Wolterink, J.M., Leiner, T., Viergever, M.A. and Išgum, I., 2018. Automatic segmentation and disease classification using cardiac cine MR images. In Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers 8 (pp. 101-110). Springer International Publishing.

[4] Le Folgoc, Loïc, Gori, Pietro. (2023) IMA205 Challenge 2023. Available at: <https://kaggle.com/competitions/ima205-challenge-2023>. Accessed: May 07, 2023.

[5] Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G. and Sanroma, G., 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?. IEEE transactions on medical imaging, 37(11), pp.2514-2525.