Multinomial Logistic Regression for Arrhythmia Detection

Omar Behadada[1], Marcello Trovati[2], Chikh MA[1,], Nik Bessis[2], **Yannis Korkontzelos**[2]

[1] Tlemcen University, Algeria
[2] **Edge Hill University, UK**

email: Yannis.Korkontzelos@edgehill.ac.uk

Edge Hill University

Summary

- We classify 48683 heartbeats of 29 patients into 4 classes
  - normal, arterial contraction, junctional contraction and ventricular contraction beats
- We adapt Multinomial Logistic Regression (MLR) as a classifier of cardiac arrhythmia
  - It learns the posterior probability distributions of each class
  - It is applied to arrhythmia detections

Medical Context & Motivation

- Electrocardiogram (ECG):
  - reflects activity of the central blood circulatory system.
  - provides information on the normal or pathological physiology of heart activity.
  - is an important non-invasive clinical tool for the diagnosis of **heart diseases**.
- Early and quick detection and classification of ECG arrhythmia is important, especially for the treatment of patients in the intensive care.

Medical Context

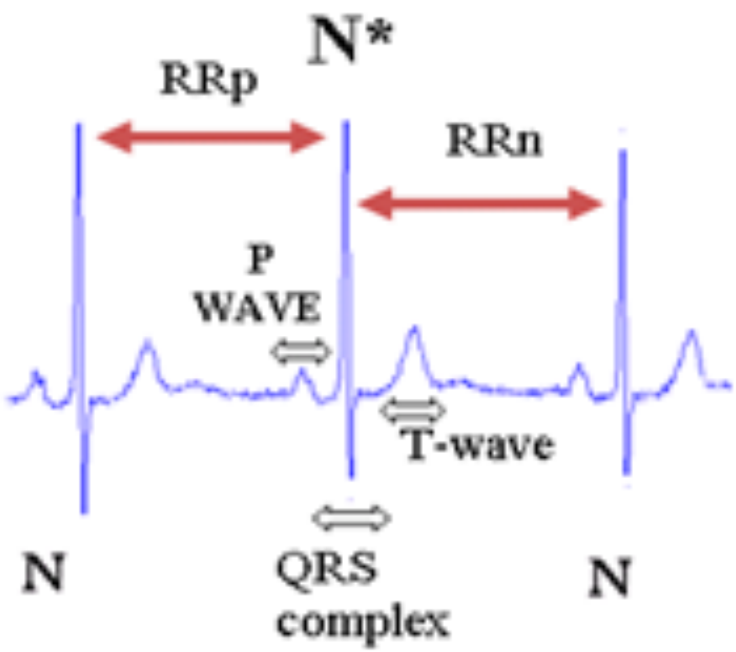
- Computer aided diagnostic (CAD) systems have been used for ECG classification.
- Popular techniques:
  - Multivariate statistics, decision trees, fuzzy logic, expert systems and hybrid approaches
- The most important step is the integration of suitable **feature extractor** and **pattern classifier**.

Data

- 29 patients of the MIT-BIH db. Conditions:
  - ventricular contraction beats (PVC)
  - premature arterial contraction beats (PAC)
  - premature junctional contraction beats (PJC)
- 48683 heartbeat samples

Beats			
N: normal	A: arterial contraction	J: junctional contraction	V: ventricular contraction
43943	711	32	3997
90.26%	1.46%	0.07%	8.21%

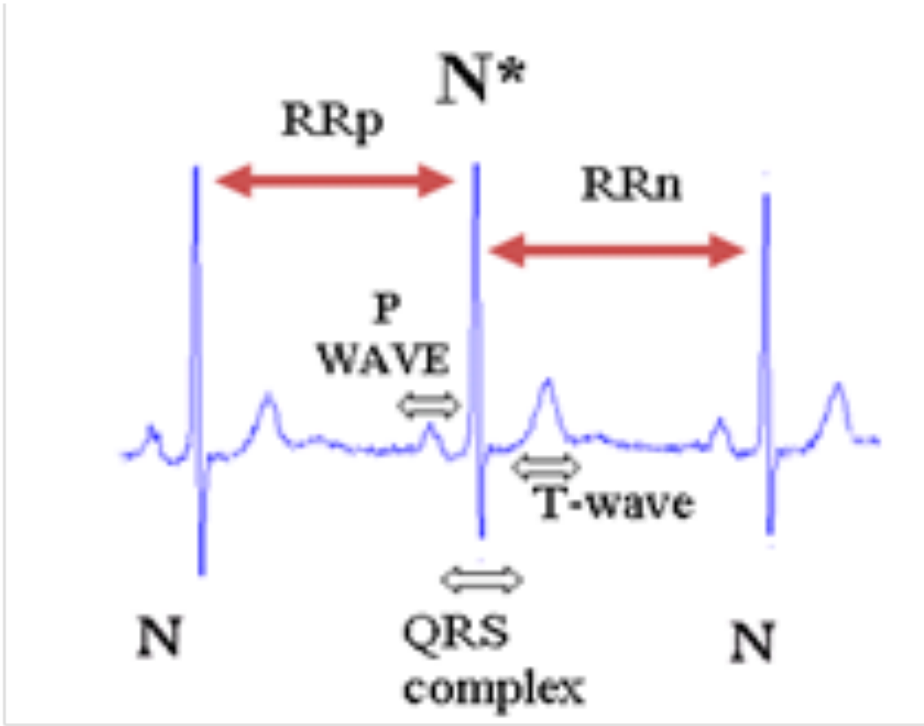
Example of ECG Beat



Attributes	Meaning
RR precedent: RR0	Distance between the peak of the current beat R and the previous one
RR next : RRn	Distance between the peak of the present R and the next beat
QRS complex	Beginning of the Q-wave and the end of the S wave
Comp	The ratio RR0/RRs
PP	Peak to peak of the R wave of the QRS complex
Energy	Energy of the QRS complex

Data Preparation

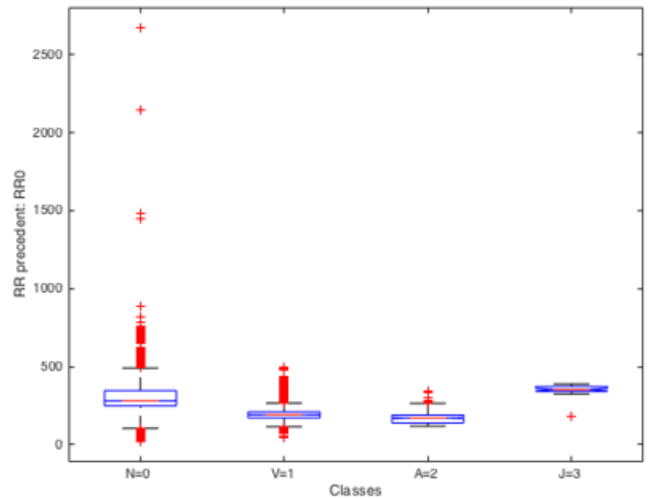
- R peaks were identified by the Tompkins algorithm

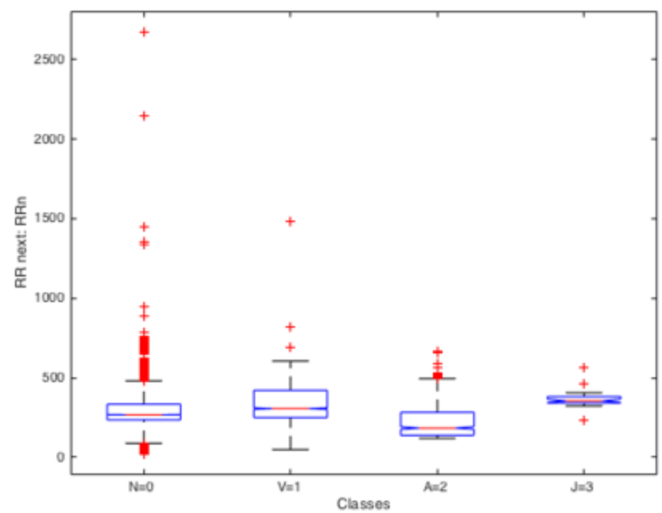


A non-trivial classification task

- Classifying heartbeats with only one feature, e.g. RR0, RRS, QRS, etc. is complex due to the lack of specific threshold for each class.

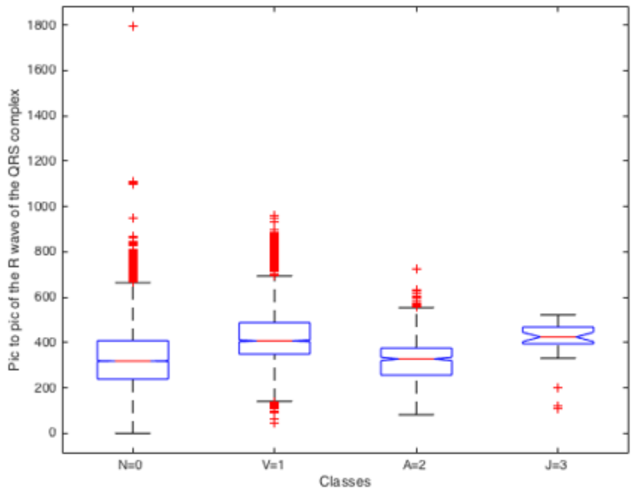
Distribution of *RRo* and *RRn*

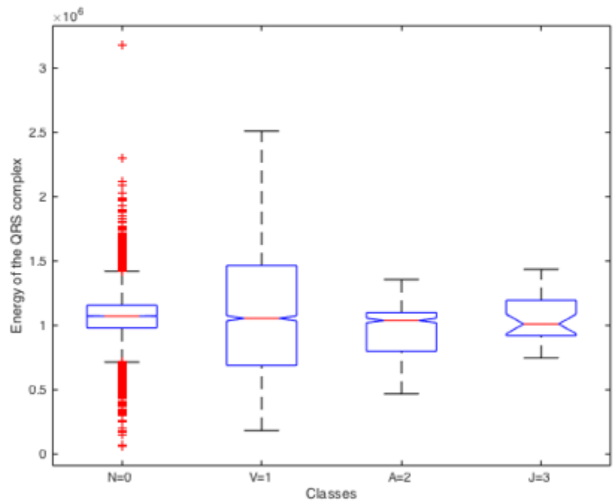




Edge Hill University

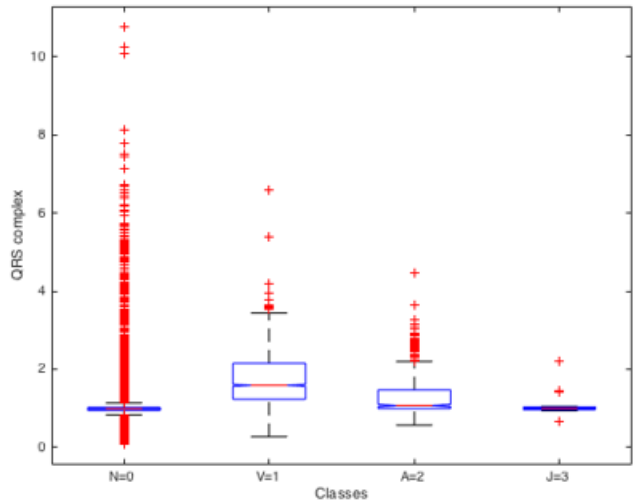
Distribution of features PP and ENERGY of QRS complex in different classes

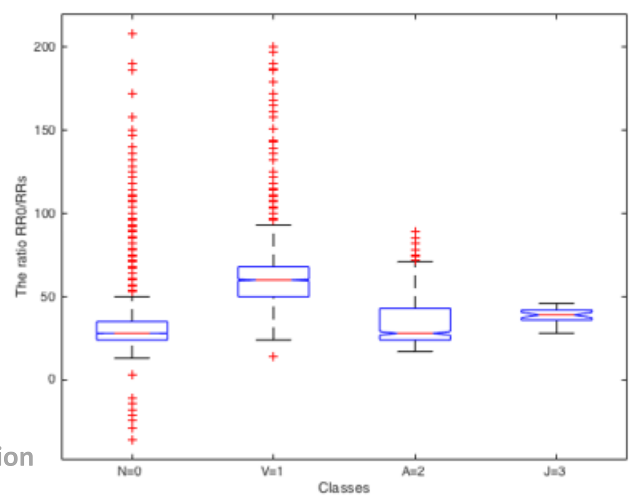




Edge Hill University

The distribution of features QRS and COMP in different classes





- We can separate between *V* and *N* classes
- This is exploited in our approach

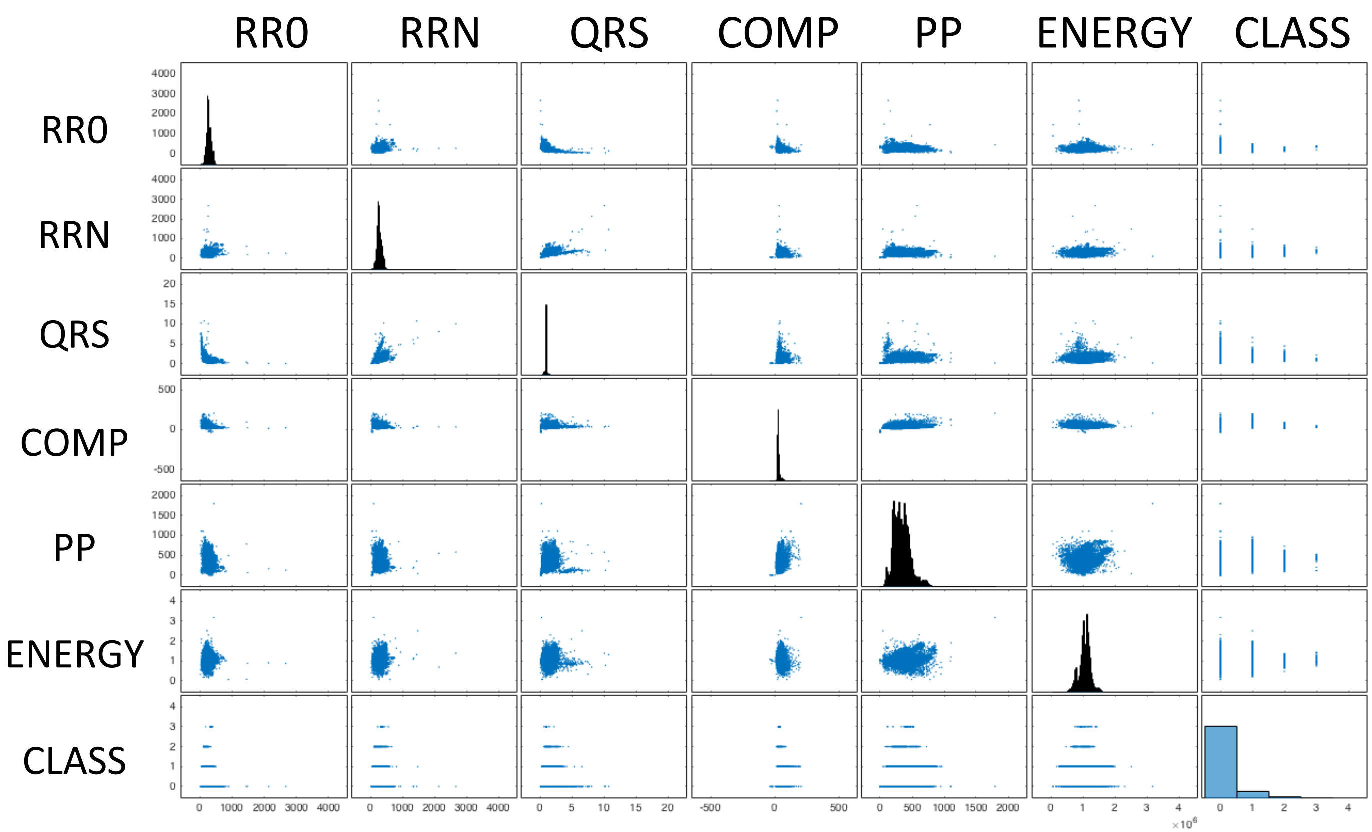
Edge Hill University

Correlation

	RR0	RRN	QRS	COMP	PP	ENERGY	CLASS
RR0	1	49.7%	43.9%	28.4%	3.7%	9.1%	36.6%
RRN	49.7%	1	41.5%	1.9%	5.7%	3.5%	4.7%
QRS	43.9%	41.5%	1	33.0%	7.6%	6.1%	**41.1%**
COMP	28.4%	1.9%	33.0%	1	17.3%	5.5%	**49.4%**
PP	3.7%	5.7%	7.6%	17.3%	1	18.6%	13.4%
ENERGY	9.1%	3.5%	6.1%	5.5%	18.6%	1	0.3%
CLASS	36.6%	4.7%	**41.1%**	**49.4%**	13.4%	0.3%	1

- QRS and COMP exhibit maximum relationships within the corresponding classes.
- This confirms that QRS and COMP are very important in arrhythmias detection.

Correlation



The histograms indicate the presence of a linear correlation.

Features

- ***RR0**: R-R* interval of the beat
  - the difference between the QRS peak of the present and previous beat
- **RRS**: ratio *RR1-to-RRo*
  - the ratio of the present over the previous R-R interval
- **QRS width**
  - calculated according to the Tompkins algorithm

- Each beat is represented as 3-dimensional vector.

Multinomial Logistic Regression (MLR)

- a supervised-learning algorithm
- a classifier able to distinguish among *K* classes
- Inputs the feature vectors of *L* labelled training samples: $D_L = \{(X_1, Y_1), \ldots, (X_L, Y_L)\}$, which is called the training set.
- Computes the posterior class distribution using to estimate regression coefficients *w*.

Logistic Regression – Training phase

The general MLR model is computed as:

$$P(y_1 = k|x_i, w) = \frac{\exp\left(w^{(k)} x_i\right)}{\sum_{k=1}^{K} \exp\left(w^{(k)} x_i\right)}$$

- w is defined as $(w^{(1)}, \ldots, w^{(K-1)})$
- $w^{(k)}$ is the set of logistic regression coefficient for class k
- $x = (x_1 , . . . , x_i)$ are the feature vectors of training samples.

Logistic Regression - Kernel

- A Gaussian Radial Basis Function (RBF) is defined as:

$$K(x_i, x_j) = \exp\left(\frac{-||x_i - x_j||^2}{2\sigma^2}\right)$$

- It describes the training vectors and offers improved data separability in the transformed space.

Logistic Regression – Training phase

- The posterior probability density of *w* with
  - Y_L: set of labels
  - X_L: set of vectors of labelled samples

$$P(w|Y_L, X_L)\alpha p(Y_L|X_L, w)p(w|X_L)$$

Logistic Regression - Coefficients

Expectation Maximization (EM) is used to estimate the regression coefficients:

$$\hat{w} = \arg max(l(w) + \log p(w|X_L))$$

where the log-likelihood function of *w* is computed as:

$$\begin{aligned} l(w) &\equiv \log p(Y_L|X_L, w \equiv \log \prod_{i=1}^{L} P(y_i|x_i, w) \\ &\equiv \sum_{i=1}^{L} x_i^T w^{(y_i)} - \log \sum_{j=1}^{K} \exp\left(x_i^T w^{(j)}\right). \end{aligned}$$

Logistic Regression – Test phase

- Regression coefficients *(w)* are constant value inputs.
- Posterior class probability densities of each feature vector are computed.
- The class label of each feature vector is determined by the index of the maximum posterior class probability.

Evaluation – Confusion Matrix

	N	V	A	J	Other
N	42016	7	9	3	0
V	16	3947	3	0	0
A	4	12	692	0	0
J	0	0	3	29	0
Other	0	0	0	1	0

- Columns represent heart beats in estimated classes, while rows represent beats in real classes.

Evaluation

Metric	Score
Last Correct Rate	93.13%
Last Error Rate	6.87%
Inconclusive Rate	0.00%
Classified Rate	100.00%
Sensitivity	92.86%
Specificity	94.17%
Positive Predictive Value	81.25%
Negative Predictive Value	97.98%
Positive Likelihood	15.94%

Evaluation: Parameters

- **CorrectRate**: Correctly Classified Samples / Classified Samples
- **ErrorRate**: Incorrectly Classified Samples / Classified Samples
- **LastCorrectRate**: CorrectRate computed only during the last results' update.
- **LastErrorRate**: ErrorRate computed only during the last results' update.
- **InconclusiveRate**: Nonclassified Samples / Total Number of Samples
- **ClassifiedRate**: Classified Samples / Total Number of Samples
- **Sensitivity**: Correctly Classified Positive Samples / True Positive Samples
- **Specificity**: Correctly Classified Negative Samples / True Negative Samples
- **PositivePredictiveValue**: Correctly Classified Positive Samples / Positive Classified Samples
- **NegativePredictiveValue**: Correctly Classified Negative Samples / Negative Classified Samples
- **PositiveLikelihood**: Sensitivity / (1 - Specificity)
- **NegativeLikelihood**: (1 - Sensitivity) / Specificity
- **Prevalence**: True Positive Samples / Total Number of Samples.

Conclusions

- Medical decision making requires automatic diagnostic support.
- Our method provides explicit knowledge base on class probabilities estimated on a medical database.
- MLR improves the transparency and interpretability of the classification process.
- In the future, we aim to integrate fuzzy partition rules into this method.

Thank you! Any questions?