DOI: 10.15514/ISPRAS-2025-37(2)-21

A Model for Atrial Fibrillation Detection Based on Differentiation and Compression of Interbeat Interval Sequences

N.S. Markov, ORCID: 0000-0001-8913-9962 <ns.markov@iip.uran.ru>

Institute of Immunology and Physiology of the Ural Branch of the Russian Academy of Sciences, 106, Pervomayskaya st., Yekaterinburg, 620078, Russia.
Ural Federal University, 19, Mira st., Yekaterinburg, 620062, Russia.

Abstract. Atrial fibrillation is the most common arrhythmia with a major impact on public health. This paper presents a model for automatic detection of atrial fibrillation episodes in ECG, using information compression and numerical differentiation for classification of beat-to-beat interval sequences. The core of the model is normalized compression distance based on the theory of universal similarity metrics. To enable class discrimination by compression we consider finite-difference representation of interval sequences with subsequent quantization procedure. In particular, we introduce a simple Δ5RR-interval representation which improves the sensitivity of the model to heart rhythm fluctuations. Our model achieves 96.37% sensitivity, 97.74% specificity and 0.935 MCC in 8x5-fold cross-validation on the MIT-BIH AFDB dataset using a segment window of 128 R-peaks. The particular advantage of the model is the classification quality in a few-shot learning setting, i.e., a training set with a small number of sequence observations can be used for classification of sufficiently large test sets.

Keywords: normalized compression distance; few-shot learning; atrial fibrillation detection; heart rate; RR-interval sequences.

For citation: Markov N.S. A Model for Atrial Fibrillation Detection Based on Differentiation and Compression of Interbeat Interval Sequences. Trudy ISP RAN/Proc. ISP RAS, vol. 37, issue 2, 2025, pp. 281-300. DOI: 10.15514/ISPRAS-2025-37(2)-21.

Acknowledgements. The author would like to thank Olga Eduardovna Solovyova for scientific supervision. The work was supported by the state task to IIP UrB RAS № 122022200089-4.

Модель детекции фибрилляции предсердий, основанная на дифференцировании и сжатии интервалограмм

Н.С. Марков, ORCID: 0000-0001-8913-9962 <ns.markov@iip.uran.ru>

Институт иммунологии и физиологии УрО РАН,
Россия, 620078, г. Екатеринбург, ул. Первомайская, 106.
Уральский Федеральный Университет имени Б.Н. Ельцина,
Россия, 620062, Свердловская область, г. Екатеринбург, ул. Мира, д. 19.

Аннотация. Фибрилляция предсердий – это наиболее распространенная в популяции аритмия, оказывающая существенное влияние на систему здравоохранения. В данной работе представлена модель автоматической детекции эпизодов фибрилляции предсердий на ЭКГ, использующая сжатие информации и численное дифференцирование для классификации последовательностей интервалов между сердцебиениями. В основе модели лежит нормализованное расстояние сжатия, основанное на теории универсальных метрик информационной близости. Чтобы обеспечить дискриминацию классов путем сжатия, в работе рассматривается конечно-разностное представление интервальных последовательностей с последующей процедурой квантования. В частности, вводится простое Δ5RR-интервальное представление последовательности, которое улучшает чувствительность модели к флуктуациям сердечного ритма. Предлагаемая модель достигает 96.37% чувствительности, 97.74% специфичности и 0.935 коэффициента корреляции Мэтьюса при 8x5-кратной кросс-валидации на базе данных MIT-BIH AFDB с использованием окна из 128 R-пиков. Особым преимуществом модели является качество классификации при обучении с малым количеством проб, то есть обучающая выборка с небольшим числом наблюдений последовательностей может использоваться для классификации достаточно больших тестовых выборок.

Ключевые слова: нормализованное расстояние сжатия; обучение с малым количеством проб; детекция фибрилляции предсердий; ритм сердца; RR-интервалограммы.

Для цитирования: Марков Н.С. Модель детекции фибрилляции предсердий, основанная на дифференцировании и сжатии интервалограмм. Труды ИСП РАН, том 37, вып. 2, 2025 г., стр. 281–300 (на английском языке). DOI: 10.15514/ISPRAS–2025–37(2)–21.

Благодарности: Автор выражает благодарность Соловьёвой Ольге Эдуардовне за научное руководство. Работа была выполнена при поддержке гос. темы ИИФ УрО РАН №122022200089-4.

1. Introduction

In a healthy heart, the cardiac rhythm is entirely governed by a special functional structure, the sinus node. Atrial fibrillation (AF) is a heart condition in which the atria beat spontaneously as the sinus node loses control over the heartbeat cycle. The rhythm of AF is characterized by an increased average frequency and quasi-chaotic fluctuations of heartbeats. The incidence of AF in the population of developed countries reaches 1-2%, making it the most common cardiac arrhythmia [1]. For this reason, AF is considered one of the most pressing public health problems.

Automatic detection of atrial fibrillation episodes is an important diagnostic task critical to patient monitoring. A lot of classification models designed for automatic detection of AF were published in the 21st century [2], as more ECG data had been made available through repositories such as PhysioNet [3]. The reliable approach to discriminate between a normal rhythm and an AF episode is to analyze irregularities in a sequence of beat-to-beat RR-intervals on ECG. RR-sequences are robust to noise [4] and can be recorded by inexpensive consumer devices [5].

Neural network-based models undoubtedly achieve the best results in the task of heart rhythm classification [2]. However, neural networks are computationally expensive to run, and the models themselves tend to overfit to certain data sets, leading to poor generalization across population [6]. This aspect makes it difficult to implement neural network models in remote health systems used for routine ECG monitoring and detection of AF episodes in patients with suspected disease.

Therefore, there is still a need for "lightweight" and efficient AF detection models that do not require large amount of data for training and can be implemented on low-cost low-energy devices. One such promising alternative is classification by compression, which uses heuristic information compression to highlight irregularities and patterns in the data.

This paper proposes a classification model that uses the proximity between interbeat interval sequences captured by a compression algorithm for automatic detection of AF episodes. The core of the model is normalized compression distance [7], which serves as an approximation of the incomputable universal similarity metric. However, to enable discriminative properties of compression we consider differentiation of RR-sequences, and a quantization procedure which turns scalar sequences into symbolic strings. For the former we introduce a Δ5RR-interval representation which uses a finite-difference scheme to emphasize rhythm fluctuations, leading to a better sensitivity of the model.

Statistical validation of the model is performed on the open MIT-BIH Atrial Fibrillation Database under repeating fivefold cross-validation. This approach allows us to evaluate the classification quality, select the best model configuration and compare it with arrhythmia detectors presented in other works.

The resulting classification model requires the selection of only two hyperparameters. Therefore, we further evaluate the quality of the model in a few-shot learning setting with a limited training sample. This allows us to assess the degree of generalization achieved by the model and its effectiveness for tasks with limited amounts of data.

2. Methods

2.1 Finite-difference representation of interbeat interval sequences

Let (R_i) be the sequence of time coordinates of R-peaks corresponding to heartbeats on the ECG recording. RR-interval sequence (RR_i) is a dynamic series of time intervals between adjacent R-peaks: $RR_i = R_i - R_{i-1}$

RR-sequences display a natural trend: a person's heart rate varies in response to a variety of external factors such as stress, exercise, stimulation, sleep, etc. A ΔRR-representation is introduced to eliminate the linear trend for classification [8]:

$$\Delta RR_i = RR_{i+1} - RR_i = R_{i+1} - 2R_i + R_{i-1}. \quad (1)$$

As can be seen from the formula, the ΔRR_i represents the difference between adjacent RR-intervals. This can be thought of as taking the second numerical derivative using a three-point scheme along the original sequence of R-peaks. We propose and investigate an analogous finite-difference Δ5RR-representation using a five-point scheme [9]:

$$\Delta_5 RR_i = \frac{-R_{i+2} + 16R_{i+1} - 30R_i + 16R_{i-1} - R_{i-2}}{12}. \quad (2)$$

In this paper three representations of interbeat sequence – RR, ΔRR and Δ5RR – were considered. The study was carried out with respect to a segment window M. This value indicates the number of R-peaks used for classification. In practice, a classifier using a smaller M is preferred because it allows the use of shorter ECG recordings for the detection of arrhythmia episodes. In this paper, the quality of classification was investigated for windows M = 32, 64 and 128.

2.2 Quantization

It is known that compression algorithms are designed primarily for binary and text data. Therefore, we propose a quantization procedure to transform interbeat interval sequences into symbolic strings.

Let X be a sequence of real values. A mapping function $q(x)$ is called a quantizer if it is defined for any $x \in X$ and takes a finite number Q of values $a_1, ..., a_Q$, which we call an alphabet □. We consider scalar quantization, i.e., $X \subseteq \mathbb{R}$. Quantizer can be viewed as a surjective function $q(x)$: $\mathbb{R} \to$ □.

From the surjectivity of the function $q(x)$ it follows that the quantizer assigns discrete states to the elements $x \in X$ depending on whether they belong to some non-overlapping subsets $\mathcal{B}_1, ..., \mathcal{B}_Q$:

$$q(x) = a_i \longleftrightarrow x \in \mathcal{B}_i \tag{3}$$

For scalar quantizer these subsets can be given as semi-intervals: $\mathcal{B}_1 \subseteq (-\infty, b_1]$, $\mathcal{B}_2 \subseteq (b_1, b_2]$, …, $\mathcal{B}_Q \subseteq (b_{Q-1}, -\infty)$. Thus, the quantizer $q(x)$ can be defined by selecting the boundaries of the semi-intervals $b_1, ..., b_{Q-1}$ and images $a_1, ..., a_Q$.

Compression algorithms are robust to alphabet permutations, so the images $\{a_i\}_{i=1..Q}$ can be chosen arbitrarily – in this paper they are Unicode symbols in UTF-8 encoding with positions from 1 to Q. On the other hand, the semi-interval boundaries $\{b_i\}_{i=1..Q-1}$ should be selected depending on the distribution of elements in X, the frequency of occurrence of particular values.

Let $c_1, ..., c_Q$ be the centroids of $\mathcal{B}_1, ..., \mathcal{B}_Q$. We can determine whether an element x belongs to a certain semi-interval by its nearest centroid:

$$x \in \mathcal{B}_i \longleftrightarrow i = \underset{i}{\operatorname{argmin}} |x - c_i| \tag{4}$$

Given the ordering of centroids (∀ □, □ $1 \le$ □ $<$ □ $\le$ □ : □□ $<$ □□) it is obvious that for the chosen {□□}□=1..□ the quantization boundaries are the midpoints on the numerical line between successive centroids:

$$b_i = \frac{c_{i+1} - c_i}{2}, 1 \le i \le Q - 1. \tag{5}$$

Thus, the problem of semi-interval boundary selection can be reduced to the problem of centroid optimization.

For a finite scalar sequence □ = (□1, □2, ..., □□) we select the semi-interval centroids by minimizing the quadratic error:

$$\underset{c_1,...,c_Q}{\operatorname{argmin}} \sum_{j=1}^{Q} \sum_{x_i \in \mathcal{B}_j} (x_i - c_j)^2. \tag{6}$$

The k-means++ clustering algorithm [10] deterministically finds near-optimal values of {□□}□=1..□ for a given X and alphabet size Q (assigning Q as the number of centroids k in the algorithm). Found scalar function $q(x)$ is used for the whole set quantization with X being the "flattened" training sample. Therefore, Q is the only parameter of the whole quantization procedure.

2.3 Classification model

The complete proposed model combines the finite-difference representation of the interbeat interval sequences and subsequent quantization with classification using normalized compression distance and k-nearest neighbour method. Below is a formalization of the entire model.

Let $\mathbb{R}^M$ be the event space. Finite monotone sequences of R-peak time coordinates are observations of events: $\mathbf{x} = (x_1, ..., x_M) \in \mathbb{R}^M$. Let $\mathcal{L}$ be the space of classes of events. We consider a binary classification, hence $\mathcal{L} = \{1,0\}$, where positive and negative classes 1 and 0 are episodes of AF and normal rhythm respectively.

Let $(\mathbf{x}_□, y_□)_{□=1..N}$ be a training sample of □ observation of events $\mathbf{x}_{1..□} \in$ □□□□□□ with known classes $y_{1..N} \in Y_{train}$. The complete classification procedure consists of four steps.

Step 1. The sequences $\mathbf{x}_i$ are transformed into either RR, ΔRR or Δ5RR-interval representation (Section 2.1):

$$\tilde{\mathbf{x}} = \mathrm{RR}(\mathbf{x}) = (x_2 - x_1, x_3 - x_2, ..., x_M - x_{M-1}) \quad (7)$$

$$\tilde{\mathbf{x}} = \Delta\mathrm{RR}(\mathbf{x}) = (x_3 - 2x_2 + x_1, ..., x_M - 2x_{M-1} + x_{M-2}) \quad (8)$$

$$\tilde{\mathbf{x}} = \Delta_5\mathrm{RR}(\mathbf{x}) = (\frac{-x_5 + 16x_4 - 30x_3 + 16x_2 - x_1}{12}, \ldots, \frac{-x_M + 16x_{M-1} - 30x_{M-2} + 16x_{M-3} - x_{M-4}}{12}) \quad (9)$$

If we denote $\tilde{x}_{i,j}$ as the value of the *j*-th element in the *i*-th sequence after the chosen finite-difference transformation, the training sample is a matrix $\tilde{X} = (\tilde{x}_{i,j})_{1 \le i \le N,\ 1 \le j \le m}$ where *m* is the new length of the sequence. Depending on sequence representation $m = M - 1$, $m = M - 2$ or $m = M - 4$.

Step 2 involves the scalar quantizer training. For this, $\tilde{X}$ is represented as a "flattened" one-dimensional scalar sequence $\tilde{X}_{flat} = (\tilde{x}_{(r \operatorname{div} m) + 1,\ (r \bmod m) + 1})$ with an abstract index $1 \le \square \le \square \times \square$. The quantizer $\square(\cdot)$ is found according to Section 2.2 by minimizing the expression (6). The size of the alphabet $\square$ is a hyperparameter of the complete classification model.

Step 3. Let $\hat{\mathbf{x}} = (\hat{x}_1, \ldots, \hat{x}_m) \in \mathbb{R}^m$ be the sequence undergoing classification (transformed to the corresponding finite-difference representation beforehand) and $\hat{y}$ be the sought unknown class. This step is the calculation of normalized compression distances between the classified observation and the sequences in the training sample.

We denote by $\mathbf{q}$ the vectorized version of the quantizer, i.e., $\mathbf{q}(\mathbf{x}) = (\square(\square_1), ..., \square(\square_m))$. Distances $(d_i)_{\square=1..\square}$ are calculated by the normalized compression distance formula [7], given as:

$$d_i = \frac{C(\mathbf{q}(\hat{\mathbf{x}})\mathbf{q}(\tilde{\mathbf{x}}_i)) - \min\Big(C(\mathbf{q}(\hat{\mathbf{x}})), C(\mathbf{q}(\tilde{\mathbf{x}}_i))\Big)}{\max\Big(C(\mathbf{q}(\hat{\mathbf{x}})), C(\mathbf{q}(\tilde{\mathbf{x}}_i))\Big)}, \quad (10)$$

where $\mathbf{q}(\hat{\mathbf{x}})\mathbf{q}(\tilde{\mathbf{x}}_i)$ is concatenation of the classified sequence with the *i*-th observation from the training sample, $C(\cdot)$ – symbol length of the compressed sequence we compute using gzip [11] compressor. Normalized compression distance, when used with compressors based on Lempel-Ziv algorithm such as gzip, serves as a pseudo-metric approximation of universal normalized information distance [12]. The universal information distance itself is defined through Kolmogorov's complexity and acts as the lowest theoretical limit for computable information metrics [7,13].

Step 4. is classification using the *k*-nearest neighbour method. Let $s(i)$ be a sorting index for $(d_i)_{\square=1..\square}$ satisfying the following conditions:

$$\forall i, j\ 1 \le i < j \le N:\ d_{s(i)} \le d_{s(j)}, \quad d_{s(i)} = d_{s(j)} \Longrightarrow y_{s(i)} \ge y_{s(j)}. \quad (11)$$

Here, in the case of equal distances, the second condition sets the sorting priority of known positive classes over negative ones in the training sample. Then, for binary classification, *k*-nearest neighbour method can be reduced to computing a single decision coefficient w_{knn}:

$$w_{knn} = \sum_{i=1}^{k} (-1)^{y_{s(i)}}, \quad (12)$$

and the sought class is determined according to the sign of the coefficient:

$$\hat{y} = h(\hat{\mathbf{x}}) = \begin{cases} 0, & w_{knn} > 0, \\ 1, & w_{knn} < 0. \end{cases} \quad (13)$$

Thus, the number of nearest neighbours *k* in (12) is the second hyperparameter of the complete model. Value of *k* must be an odd number to avoid choice uncertainty in (13).

Steps 3 and 4 extend element-wise to any set of classified sequences $\{\hat{\mathbf{x}}_i\}_{i=1..l}$ in the testing sample X_{test} without the need for quantizer retraining in Step 2. Since the complete model does not have hidden parameters (intermediate parameters non-deterministically calculated according to the input), the information model can be viewed as a family of classifiers $\mathscr{H}(Q, k)$. The choice of the sequence representation in Step 1 is another part of the model configuration. This paper explores the influence of selected finite-difference representation (RR, ΔRR or Δ5RR), the number of nearest neighbours *k* and the alphabet size *Q* on the quality of performed classification.

2.4 Data

The efficacy of the model was investigated using the open MIT-BIH Atrial Fibrillation Database (AFDB), one of the most popular databases for the validation of AF classification methods [14-15]. AFDB consists of 25 Holter ECG monitoring recordings of 10 hours duration, sampled at 250 Hz. The database comes annotated with time coordinates of R-peaks and class labels for signals.

The R-peak coordinate sequences and their corresponding class labels were extracted from each recording and partitioned into sub-sequences of non-overlapping windows of length *M* = 32, 64, 128. A full sample was created for each considered *M*. Subsequences included normal or AF rhythm only. If, according to the annotations, a transition from one rhythm to another was observed in an ECG corresponding to the window, that subsequence was excluded from the sample. The number of subsequences in both classes extracted for each window size is shown in Table 1. There is a slight class imbalance in the samples with a bias towards normal rhythm sequences.

In addition to AFDB, a MIT-BIH Arrhythmia Database (MITDB) was used as an external test sample [3]. MITDB consists of 48 half-hour ECG recordings sampled at 360 Hz. Eight of the contained recordings included AF episodes. Interval sequences of normal rhythm from all 48 recordings and of AF episodes from 8 recordings were extracted for each window size. Other rhythm types were excluded. Not surprisingly, the extracted normal rhythm sequences were significantly more numerous than the AF sequences (Table 1).

Table 1. Number of interbeat interval sequences per class extracted from AFDB and MITDB recordings for each window size M.

AFDB					
Extracted for *M* = 32		Extracted for *M* = 64		Extracted for *M* = 128	
Normal	AF	Normal	AF	Normal	AF
20447	16032	10106	7896	4975	3851
Overall: 36479		Overall: 18002		Overall: 8826	
MITDB					
Extracted for *M* = 32		Extracted for *M* = 64		Extracted for *M* = 128	
Normal	AF	Normal	AF	Normal	AF
2212	289	1028	124	456	49

Overall: 2501	Overall: 1152	Overall: 505

2.5 Validation procedures

Statistical validation of the classification model was performed with two data partitioning procedures: using the full available dataset and under few-shot learning setting. The procedures differed in ways the splits into training and test samples were formed, and in the sizes of the training samples.

The full dataset procedure was a repeating fivefold cross-validation on the population. AFDB recordings were divided into 5 random groups of equal size, from which 5 possible population splits were created, with one group being the test population and the other 4 groups combined to form the training population. The procedure was repeated 8 times. This resulted in 40 splits into training and test populations of 20 and 5 recordings respectively. Extracted interval sequences from recordings were divided into training and test samples according to populations in order to prevent data leaks.

The external dataset procedure was used to supplement cross-validation on a single database, Experiments using MITDB as a test sample were performed with the whole AFDB database used as a training sample.

The few-shot learning procedure was used to assess the quality of classification in the context of a limited amount of training data. AFDB recordings were randomly divided into a training and a test population of 20 and 5 recordings respectively. All extracted interval sequences from the recordings of the test population formed the test sample. In turn, n random observations of sequences of both classes were randomly selected from the 20 recordings of the training population. This procedure was repeated 500 times for n = 5, 10, 20, 50, 100. I.e., 500 experiments were performed with a training sample consisting of n normal rhythm sequences and n AF sequences for each value of n.

The splits within procedures were made for a fixed window M and sequence representation. As three window sizes (M = 32, 64, 128) and three interval sequence representations (RR, ΔRR, Δ5RR) were investigated, classifications were repeated with the resulting training-test samples for all nine window/representation combinations. The same population splits were used for each configuration.

The Matthews correlation coefficient was used as the main measure to compare the quality of classification. Given the values of confusion matrix, MCC is determined by the formula:

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (14)$$

where TP, TN, FP, FN are the numbers of true-positive, true-negative, false-positive, false-negative classifications. MCC is considered one of the most informative measures of binary classification quality [16], because it accounts for all four underlying basic rates of confusion matrix (sensitivity, specificity, predictive positive and negative values) and assigns equal importance to ratios of successful positive and negative predictions. The latter is particularly important given the class imbalance (Table 1).

In the present work, the average MCC between validation samples was the main quality measure for a fixed model configuration. In addition to MCC, sensitivity and specificity values indicating the proportions of true-positive and true-negative classifications are also provided.

3. Results

3.1 Distance matrix and quantization analysis

To make a preliminary assessment about the ability of the proposed model to discriminate between classes we selected a random training-test sample of two non-overlapping groups of patients from AFDB, each consisting of 100 observations per normal rhythm and AF rhythm (M = 128). Matrices

of compression distances (10) calculated for Δ5RR, ΔRR, RR representations with columns and rows sorted by class are presented in Fig. 1.

We can visually assess that the matrices for all three representations were clearly divided into 4 square clusters and show the necessary characteristics for classification – between-class distances (top-right and bottom-left parts of the matrices) are higher than intra-class distances (top-left and bottom-right parts of the matrices). This observation was confirmed by average distances $\bar{d}$ inside clusters and their 95% confidence intervals (provided in Fig. \ref{img:ch2distance_matrix_big}).

However, the RR representation led to the worst separability of the AF class, as the between-class distances ($\bar{d}$ = 0.757, 0.767) were not much higher than intra-class distances ($\bar{d}$ = 0.726). This can be said in contrast to Δ5RR representation which led to the highest between-class distances ($\bar{d}$ = 0.788, 0.786) and the lowest intra-class distances for the AF class ($\bar{d}$ = 0.647). In turn, the ΔRR representation led to the lowest intra-class distances for normal rhythm ($\bar{d}$ = 0.591). Based on these results, we can assume that the use of Δ5RR in the model achieves fewer false-negatives (meaning higher sensitivity) and better quality overall. The use of ΔRR results in fewer false-positive classifications and therefore higher specificity. Qualitatively, these results were observed for any random sample.

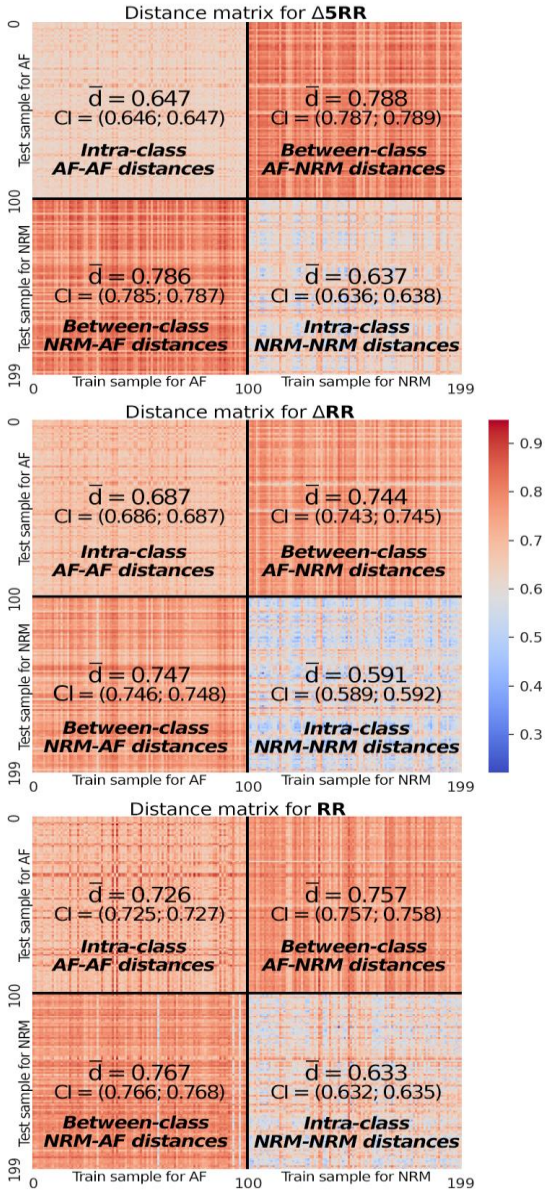

Fig. 1. Compression distance matrices for Δ5RR, ΔRR, RR between two data samples. The columns and rows of the matrix are sorted by AF and normal rhythm (denoted as NRM in the Figure). The mean distances and 95% confidence intervals (calculated by bootstrapping [17]) are provided for rectangular clusters.

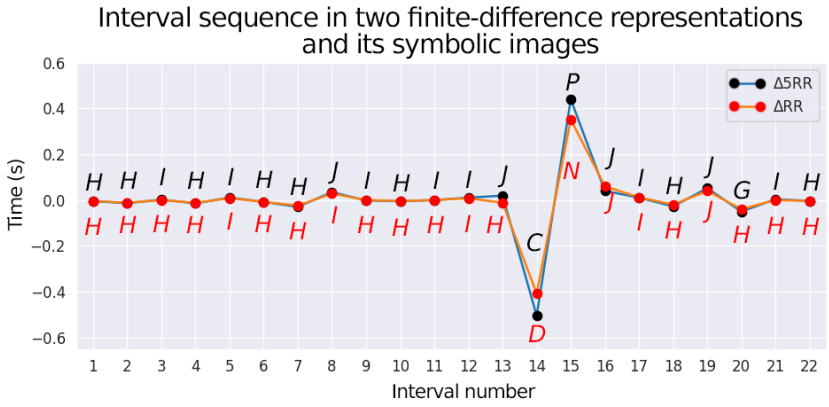

Fig. 2. Interval sequences of a normal rhythm episode with a single arrhythmic contraction, shown in Δ5RR and ΔRR representations. Symbolic images for each interval after quantization (black letters denote images for Δ5RR representation, red letters for ΔRR) are given as well.

Analysis of interval sequences and their symbolic images after quantization allows us to suggest an explanation for the changes in the distances obtained for the two finite-difference representations. Fig. 2 shows an example of a short interval sequence of the same rhythm episode with superimposed Δ5RR and ΔRR representations and their symbolic images after quantization (Q = 33). This episode corresponded to normal rhythm with a single arrhythmic contraction of the heart. The values of the intervals between two representations were quite close to each other, but the magnitudes of peaks 14 and 15 corresponding to the arrhythmic contraction were higher for five-point scheme ($\Delta5RR_{14,15}$ = (–0.504, 0.440); $\Delta RR_{14,15}$ = (–0.408, 0.352)). Positions 13 and 16 also differed between representations ($\Delta5RR_{13,16}$ = (0.019, 0.040); $\Delta RR_{13,16}$ = (-0.012, 0.060)) because the five-point finite-difference scheme is informed by two adjacent rapid rhythm changes instead of one.

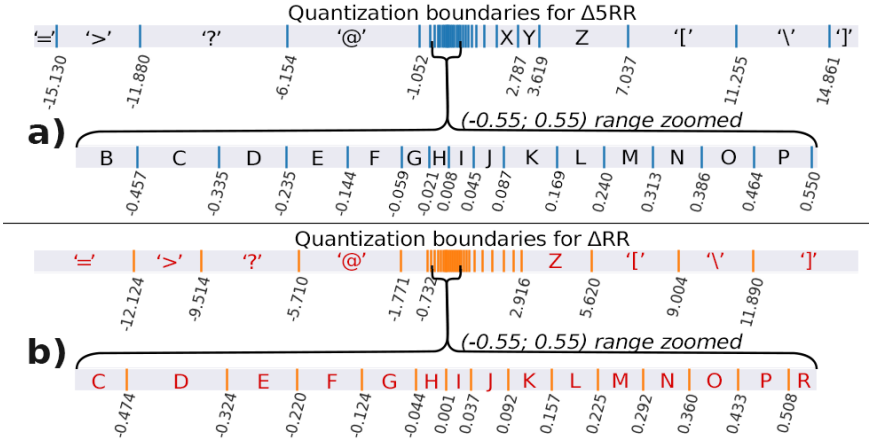

*Fig. 3. Quantization boundaries learned for Δ5RR (**a**) and ΔRR (**b**) representations with the same training sample (Q=33). Top rows of sub-figures are full boundary spreads on the same scale, bottom rows are zoomed to (-0.55; 0.55) range.*

Such changes in the interval distributions affect the resulting quantization boundaries. As can be seen in Fig. 3, the quantization boundaries for Δ5RR when compared to ΔRR were more consolidated around zero, but spread further when removed from zero. This was reflected in symbolic images of interval sequences in Fig. 2. The prefix before the peaks, reflecting a regular rhythm before an unscheduled heart contraction, appeared in the ΔRR representation as "H-H-H-H-I-H-H-I-H-H-H-I-H". This is evidently a more homogeneous string than "H-H-I-H-I-H-H-H-J-I-H-

I-I-J" in the Δ5RR representation. A similar effect was observed for the last three interval positions, appearing as "H-H-H" in ΔRR-representation and "G-I-H" in Δ5RR-representation. In other words, we can assume that the Δ5RR scheme makes it possible to obtain a quantizer that is more sensitive to the instantaneous changes in heart rhythm that are characteristic of arrhythmias. Differences in compression distances (Fig. 1) are a consequence of this effect.

3.2 Hyperparameter grid test

Since the classification model has only two hyperparameters Q and k, we conducted a grid test of classification quality. For this purpose, a full dataset validation was performed for each pair of hyperparameters from the Cartesian product of subsets Q = {30, 33, 36, ..., 114, 117} and k = {1, 3, 5, ..., 499, 501}. Results in the form of average MCC heatmaps for windows M = 32, 64, 128 using the ΔRR representation are shown in Fig. 4.

According to the obtained heatmaps, alphabet size had a negligible effect on the classification quality for small window M = 32, but larger alphabet size improves the classification quality for larger windows M = 64 and 128. The heat maps are given for the ΔRR scheme, but other representations showed a similar trend. For the remainder of the paper, Q was fixed (Table 2). For windows M = 64, 128 the same alphabet size Q = 102 was selected. For M = 32 it was fixed to Q = 39. These values were selected as the highest classification quality was achieved for them. A more detailed statistical analysis of the model classification results with respect to the number of neighbours k with fixed Q is given in the following sections.

3.3 Classification on the full dataset

In this section, we present the full dataset classification results with quality comparison between RR, ΔRR and Δ5RR sequence representations for the three investigated window sizes.

Fig. 5a shows the changes in average MCC scores and their 95% confidence intervals upon the number of nearest neighbors k for window M = 128. Confidence intervals for each representation here and below were constructed using the bootstrapping method [17], as the classification quality scores do not follow normal distribution laws. For a large window size, the Δ5RR representation significantly outperformed the ΔRR representation in terms of classification quality. The confidence intervals of the baseline RR representation overlapped with ΔRR for some optimal values of k, but generally showed a deterioration in classification quality with the increase of the number of nearest neighbors.

Analogous graphs for window M = 64 are shown in Fig. 5b. As in the previous case, the classification quality for Δ5RR was significantly higher than for ΔRR. The basic RR representation was behind both finite-difference representations.

Fig. 5c shows the MCC scores for small window M = 32. In this case we cannot confirm the advantage of one finite-difference representation over the other due to the intersection of the confidence intervals with the average curves. However, the classification quality of the baseline RR representation remained below that of the two finite-difference representations.

Presented graphs show that the MCC curves rose for small values of k, peaking for $k \leq 89$ and then entering a slow decline. Peak k values varied between different M and sequence representations. Empirically, in order to achieve high classification quality, we suppose that k can be chosen as follows:

$$k_{empiric} = \mathbf{odd}(\sqrt[3]{N}), \tag{15}$$

where N is the number of observations in the training sample, **odd**(·) is rounding to the nearest odd number.

Aggregated quality scores for classifiers with empirically derived k comparing all sequence representations in terms of MCC, sensitivity and specificity are presented in Table 3. Using Δ5RR

scheme instead of ΔRR resulted in significantly better classification quality for medium and large window lengths (M = 64, 128), as confirmed by the averages and confidence intervals of MCC. The classification model performed best for window M = 128, where Δ5RR achieved 96.37% sensitivity and 97.74% specificity, and ΔRR achieved 87.78% sensitivity and 99.00% specificity. The use of ΔRR representation led to higher classification specificity at the expense of sensitivity. This result confirms the assumptions about the sensitivity and specificity of the classification model made in Section 3.1 based on the analysis of the distance matrices (Fig 1). This means that the use of a five-point finite difference scheme can reduce the number of false-negative predictions, and therefore provide a more pathology-sensitive model.

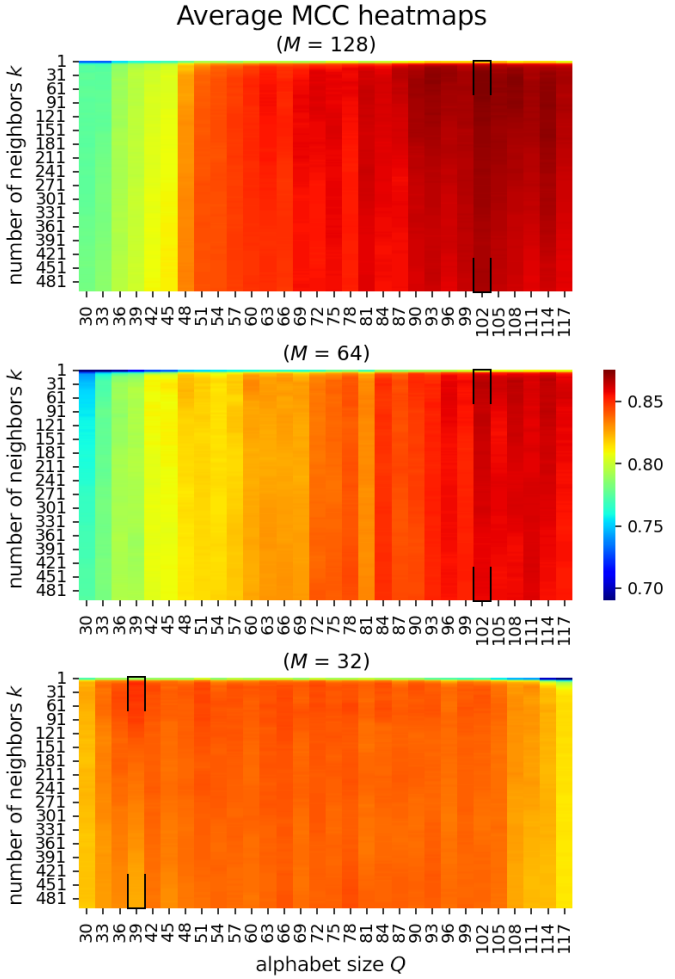

Fig. 4. Heatmaps of average MCC scores using full dataset validation on a k × Q hyperparameter grid. The maps are presented for three investigated windows M with the fixed ΔRR scheme.

Table 2. Fixed values of parameter Q relative to window size M.

Window *M*	Fixed *Q*
32	39
64	102
128	102

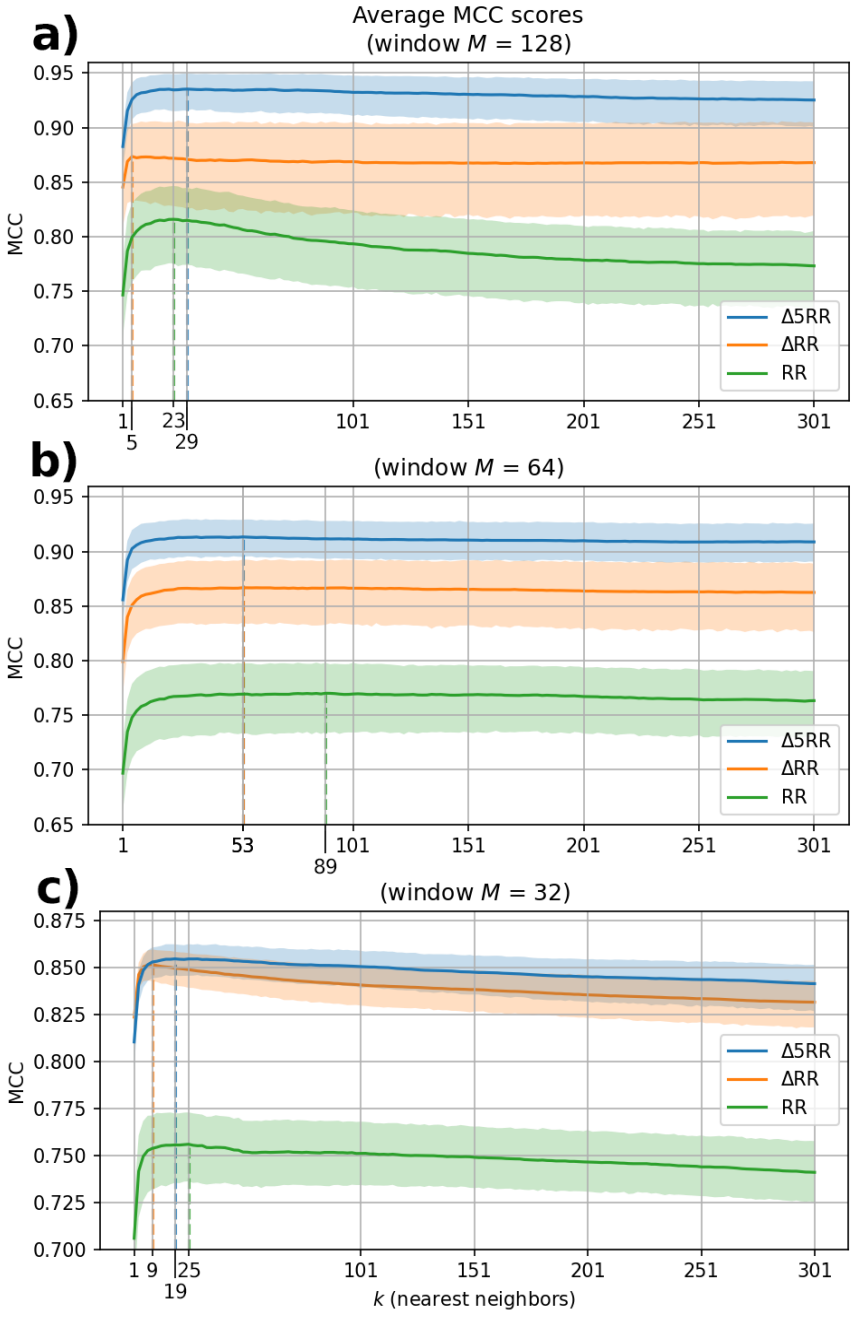

*Fig. 5. Graphs showing dependence of average MCC scores on the number of neighbours k per sequence representation. 95% confidence intervals were calculated using the bootstrapping method [17]. Dashed lines indicate the highest scores achieved for each of the three representations. Subfigures **a**, **b** and **c** correspond to windows M = 128, 64 and 32 respectively.*

The above results allow us to conclude that both finite-difference representations significantly outperform the baseline RR representation in terms of classification quality for all considered M.

Therefore, in the next section on the few-shot learning validation procedure, the results of only ΔRR and Δ5RR representations are reported as clearly preferred approaches.

Table 3. Classifiers in terms of average MCC, sensitivity and specificity with empirically derived values of k in the full dataset 8×5-fold cross-validation setting.

Δ5RR						
M	$k_{empiric}$	CI-	Average MCC	CI+	Average sensitivity	Average specificity
32	k=31	0.845	0.854	0.862	93.89%	91.87%
64	k=25	0.895	0.913	0.929	95.05%	96.84%
128	k=19	0.916	0.935	0.949	96.37%	97.74%
ΔRR						
M	$k_{empiric}$	CI-	Average MCC	CI+	Average sensitivity	Average specificity
32	k=31	0.838	0.848	0.857	94.90%	90.36%
64	k=25	0.833	0.866	0.890	88.55%	97.94%
128	k=19	0.829	0.874	0.906	87.78%	99.00%
RR						
M	$k_{empiric}$	CI-	Average MCC	CI+	Average sensitivity	Average specificity
32	k=31	0.736	0.755	0.772	80.68%	93.52%
64	k=25	0.729	0.767	0.796	79.28%	96.53%
128	k=19	0.773	0.815	0.845	82.29%	97.87%

3.4 Classification on the external dataset

This section presents results of external dataset classification. MITDB was the classified test sample, AFDB was the training sample. Hyperparameters *Q* and *k* were selected according to Table 2 and expression (15) in preceding sections. We conducted the experiments with both finite-difference representations and three window sizes. Results are provided in Table 4.

A satisfactory result was observed for the medium window $M = 64$ with Δ5RR scheme leading to 96.77% sensitivity, 96.00% specificity and 0.829 MCC. The best classification quality was achieved for the large window $M = 128$, with similar specificity (98.46% and 96.71%) and sensitivity (both 97.96%) between Δ5RR and ΔRR. However, even a small increase in the number of false-positive predictions with ΔRR (reflected in specificity) caused a considerable reduction of the MCC score (0.916 against 0.840 for $M = 128$). This was due to a significantly uneven class distribution in MITDB (see Table 1).

The impact of imbalanced classes was especially evident in the case of $M = 32$. Classification of smaller interval sequences led to a significant increase in false positives with a marked effect on MCC. Sensitivity was quite high, but 87.45% and 83.45% specificity for Δ5RR and ΔRR representations resulted in 0.646 and 0.591 MCC scores.

Table 4. Classifiers in terms of average MCC, sensitivity for external dataset classification.

Δ5RR			
M	MCC	Sensitivity	Specificity
32	0.646	96.54%	87.45%
64	0.829	96.77%	96.00%
128	0.916	97.96%	98.46%
ΔRR			
M	MCC	Sensitivity	Specificity
32	0.591	97.58%	83.45%
64	0.757	98.39%	93.00%
128	0.840	97.96%	96.71%

Outlined test results provide further evidence of the proposed model's high classification quality, especially for medium and large sequence windows. For short windows $M = 32$ the quality may be further improved with additional hyperparameter tuning. Curiously, in contrast to the previous section, the use of the Δ5RR scheme improved the specificity of the classification rather than the sensitivity. This may be due to differences in the datasets, as subjects in MITDB were affected by a broader range of heart rhythm conditions, while in AFDB the majority were AF only.

3.5 Few-shot learning

This subsection presents classification results under the limited dataset condition where training samples consisted of n observations per two rhythm classes. The few-shot learning validation was performed for $n = 5, 10, 20, 50, 100$. The number of nearest neighbors for each n was fixed at $k = \mathbf{odd}(\sqrt{(2 \times n)})$ (square root instead of cubic root in (15)). ΔRR and Δ5RR representations were compared between three investigated window sizes. As all training samples corresponded to the same rhythm episodes in different interval representations, the Wilcoxon test with Benjamini-Hochberg correction [18] was used to assess the significance of differences between ΔRR and Δ5RR ($\alpha = 0.05$).

For the window $M = 128$, the classification results are shown as box plots in Fig. 6a. One can see that the median MCC values were quite high, but for $n = 5$ there were significant outliers down to MCC = 0. The latter indicates the occurrence of small samples leading to completely incorrect classifications. For each value of n, there was a significant advantage of the Δ5RR representation over ΔRR, confirmed by the Wilcoxon criterion with multiple test correction (all p-values < 0.001).

In turn, for window M = 64, classification results are presented in Fig. 6b. The use of shorter sequences for classification led to an increase in the spread of MCC scores and a higher number of outliers for $n = 5,10$, which reduced the classification quality. The advantage of the Δ5RR representation for classification was still observed (highest $p = 0.027$ for $n = 5$).

For window $M = 32$, results are shown in Fig. 6c. Similar to the classification on the full dataset, there was no clear advantage of the Δ5RR scheme over ΔRR when using short sequences. In this

case, for $n = 5$ the Wilcoxon test did not show a statistically significant difference ($p = 0.138$). For $n = 5, 10$, there was also a wide spread of MCC scores and a large number of outliers.

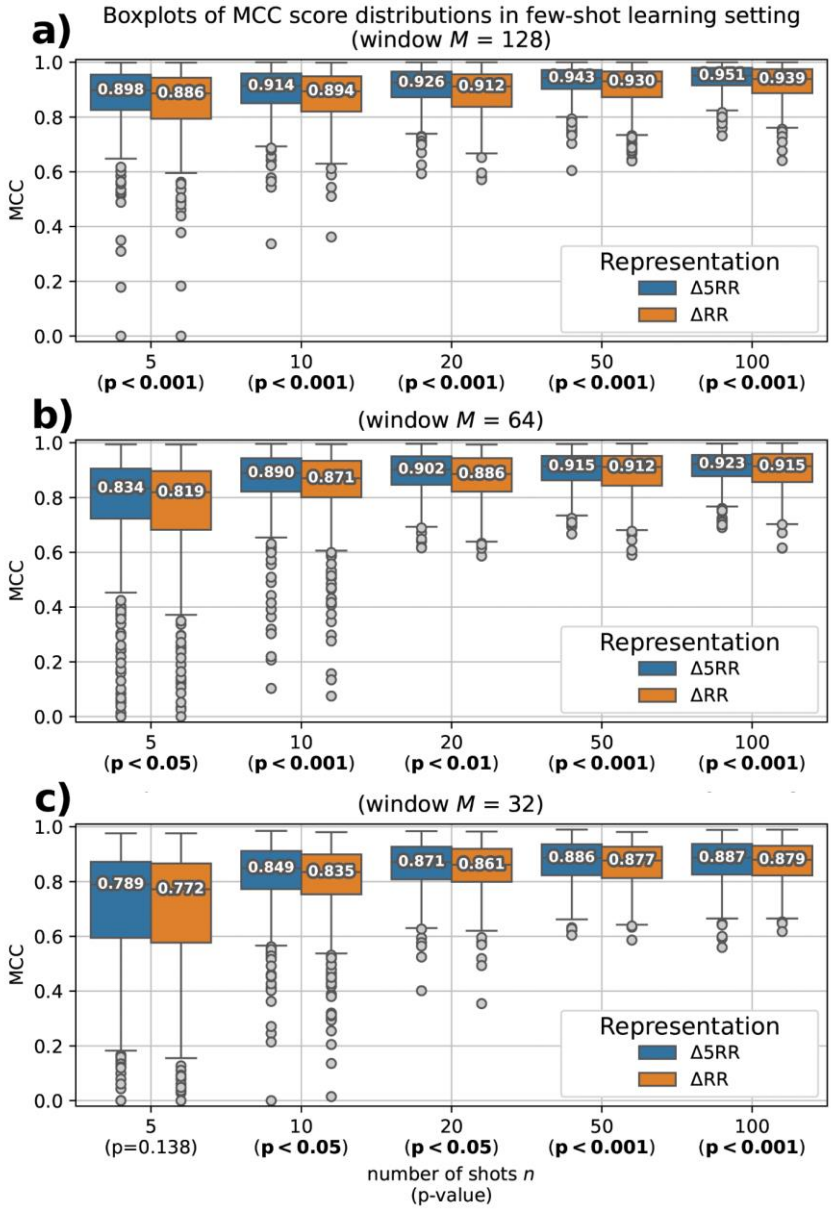

Fig. 6. Boxplots showing distributions of MCC scores in the few-shot learning setting for varying shots n. Numbers on boxes are MCC medians, boxes represent the interquartile range, whiskers show the edges of significant values within the radius of IQR × 1.5, grey dots show outliers.
Also presented are multiple test adjusted p-values of the Wilcoxon criterion for each n.
*Subfigures **a**, **b** and **c** correspond to windows M = 128, 64 and 32 respectively.*

Based on the results presented, we can say that the proposed classification model has a high generalization capability. For medium and large windows ($M = 64, 128$), the Δ5RR scheme allows a significant increase of the classification quality. Requirement to the number of samples n for stable classification depends on the window size. For M = 32, a minimum of 20 observations of the two

rhythm classes is preferred, which was ≈ 0.0011% of the full sample (Table 1), achieving a median MCC of 0.871. For M = 128, at least 10 observations of each rhythm are preferred for stable classification, which was ≈ 0.0023 of the full sample (Table 1), achieving a median MCC of 0.914.

In general, the model demonstrates high classification quality in the few-shot learning setting, but increasing the window *M* and expanding the training sample leads to better results. It is worth noting that the small training samples were randomly selected from interbeat interval sequences of 20 patients, while the test samples consisted of the full set of sequences from 5 patients. This means that a good practice for classification with a limited data set is to use observations from different patients.

4. Discussion and Conclusions

In this paper we proposed the classification model using information compression and numerical differentiation of interbeat interval sequences to detect AF on ECG. A finite-difference scheme which improves classification by compression when applied to interval sequences was proposed. We explored the choice of hyperparameters *k* and *Q*, three representations of interbeat interval sequences, the size *M* of the R-peak segment window.

As a method for arrhythmia detection, the model achieves strong results. With a segment window size of *M* = 64 (corresponding to 45-60 seconds of ECG), the model achieved a mean sensitivity of 95.05% and a mean specificity of 96.84% with 8×5-fold cross-validation on the full AFDB set. With a wider window size of *M* = 128 (90-120 seconds ECG), the model achieved a mean sensitivity of 96.37% and a mean specificity of 97.74%. For a short window of *M* = 32 (22-30 seconds ECG), the model achieved a mean sensitivity of 93.89% and a mean specificity of 91.87%. Using MITDB as the external test set and AFDB as the training set the model achieved 97.96% sensitivity and 98.46% specificity for the window *M* = 128. However, the main advantage of the method lies in the quality of classification in the few-shot learning setting, where 10-20 observations of two rhythm types are sufficient to classify entire test samples.

The results obtained on the open MIT-BIH AFDB database let us compare the proposed model with published works. The compression-based classification has the following advantages over other AF detection methods:

- Tateno's model [8] was proposed in one of the key studies in the field. This paper proposed the use of ΔRR-interval sequences for classification. The methodology involved histogram-template matching. In this study, 94.4% sensitivity and 97.2% specificity were achieved for a window size of *M* = 100 with AFDB and MITDB being used as respective training and test samples. However, the classification performance is significantly reduced for shorter sequences. In our work, we proposed a Δ5RR-interval representation, which improved the classification quality compared to ΔRR.
- Andersen's work [19] explores the possibility of AF classification using a support vector machine model. The authors claimed 96.8% sensitivity and 96% specificity for window *M* = 100 with fivefold cross-validation on AFDB (but without multiple validation repeats, unlike our work). Fivefold cross-validation purports 80-20 train-test split and the authors did not investigate classification using a more conservative data division. Presumably this is due to kernel models such as support vector machine requiring a significant amount of data for training and generalization.
- Neural network classifiers such as Xia's [20] show some of the best results in terms of quality (98.7% sensitivity, 98.9% specificity). However, neural network models require significant computational resources, and are prone to overfitting [6].

The few-shot learning paradigm originated in the field of neural networks, where classification models have achieved widespread success with large amounts of data, but struggled in applications with small available datasets. A significant amount of published work in this area has focused on

improving pre-trained models with available prior knowledge [21]. From rhythm detection applications, a transfer learning of a large pre-trained neural network model was conducted in work [22] using a small number of ECG recordings of 30-180 seconds duration for personalized AF classification. The great advantage of the model proposed in this paper is that it requires no prior information: the hypothesis space is completely determined by the used training sample. Therefore, it may be promising to use our approach to build truly personalized classifiers.

During the model validation and testing of different sequence representations, it was shown that the numerical Δ5RR scheme introduced in our work outperforms ΔRR in most cases (in particular for windows $M = 64, 128$). The result of the Δ5RR representation can be explained by the higher order of accuracy of the five-point scheme compared to the three-point scheme. This gives more emphasis to sudden rhythm fluctuations and leads to a more sensitive sequence quantizer (Fig. 2, 3).

The disadvantage of the proposed classification model is the computational demand when dealing with large amounts of data. To classify a sample of size l using a training sample of size n, it is necessary to compute a distance matrix of $l \times n$ dimensions, which involves compression of all individual sequences and all pairwise concatenations. This leads to a time complexity of at least $O(l \times n + n + l)$. That is why, from a practical point of view, it is of utmost importance that the presented model demonstrates high classification performance on small training samples.

It should be noted that gzip compression is often optimized at the hardware level [23]. This, together with the model's low data and memory requirements, makes classification with compression particularly promising for low-energy medical devices: ECG Holter monitors, portable electronics, telemedicine and remote patient monitoring. Due to proposed model's generalization capability, its practical healthcare use could be preferable to “compressed” neural network classifiers of heart rhythms. Such models use a reduced number of hidden parameters and layers to lower complexity and increase the inference performance, making them more suitable for wearable devices [24]. However, it is currently postulated that neural networks with higher number of hidden parameters lead to better generalization [25]. Introducing shallow networks into medical practice may require their costly retraining to each local population. In contrast, our model may be more easily adapted to population shift with a few rhythm observations.

It is of further interest to extend the model to multi-class classification. Atrial fibrillation is one of the most common heart diseases and it is important to investigate detection of rarer arrhythmias. Since normalized compression distance was initially proposed for clustering tasks [12], there is another prospect of non-supervised learning without labeling cardiac interval sequences into rhythms.

Normalized compression distance was originally invented for time series of discrete states, such as texts and discrete sequences [7]. In turn, the classification model proposed in this paper is suitable for monotonic stochastic sequences. The methodology considered, involving differentiation of sequences, quantization and compression distances, was also used by the author to classify subjects with dyslexia using eye-tracking data [26]. Thus, we can expect further results of the present model in other applications.

References

[1]. Staerk L., Sherer J. A., Ko D., Benjamin E. J., Helm R. H. Atrial fibrillation: epidemiology, pathophysiology, and clinical outcomes. Circulation research, vol. 120, issue 9, 2017, pp. 1501-1517. DOI:10.1161/CIRCRESAHA.117.309732.

[2]. Faust O., Ciaccio E. J., Acharya U. R. A review of atrial fibrillation detection methods as a service. International journal of environmental research and public health, vol. 17, issue 9, 2020, p. 3093. DOI:10.3390/ijerph17093093.

[3]. Moody G. B., Mark R. G. The impact of the MIT-BIH arrhythmia database. IEEE engineering in medicine and biology magazine, vol. 20, issue 3, 2001, pp. 45-50. DOI:10.1109/51.932724.

[4]. D'Aloia M., Longo A., Rizzi M. Noisy ECG signal analysis for automatic peak detection. Information, vol. 10, issue 2, 2019, p. 35.

[5]. Georgiou K., Larentzakis A. V., Khamis N. N., Alsuhaibani G. I., Alaska Y. A., Giallafos E. J. Can wearable devices accurately measure heart rate variability? A systematic review. Folia medica, vol. 60, issue 1, 2018, pp. 7-20. DOI:10.2478/folmed-2018-0012.

[6]. Zhou K., Liu Z., Qiao Y., Xiang T., Loy C. C. Domain generalization: A survey. IEEE transactions on Pattern Analysis and Machine Intelligence, vol. 45, issue 4, 2004, pp. 4396-4415. DOI:10.1109/TPAMI.2022.3195549.

[7]. Li M., Chen X., Li X., Ma B., Vitányi P. M. B. The similarity metric. IEEE transactions on Information Theory, vol. 50, issue 12, 2004, pp. 3250-3264. DOI:10.1109/TIT.2004.838101.

[8]. Tateno K., Glass L. Automatic detection of atrial fibrillation using the coefficient of variation and density histograms of RR and ΔRR intervals. Medical and Biological Engineering and Computing, vol. 39, 2001, pp. 664-671. DOI:10.1007/BF02345439.

[9]. Fornberg B. Generation of finite difference formulas on arbitrarily spaced grids. Mathematics of computation, vol. 51, issue 184, 2017, pp. 699-706. DOI:10.1090/S0025-5718-1988-0935077-0.

[10]. David A., Sergei V. k-means++: The Advantages of Careful Seeding. In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, 2007, pp. 1027-1035. Available at the URL: http://ilpubs.stanford.edu:8090/778/.

[11]. Gailly J.L., Adler M. GNU gzip. GNU Operating System, 1992. Available at the URL: https://www.gnu.org/software/gzip/manual/gzip.pdf.

[12]. Cilibrasi R., Vitányi P. M. B. Clustering by compression. IEEE Transactions on Information theory, vol. 51, issue 4, 2005, pp. 1523-1545. DOI:10.1109/TIT.2005.844059.

[13]. Bennett C. H., Gács P., Li M., Vitányi P. M. B., Zurek W. H. Information distance. IEEE Transactions on information theory, vol. 44, issue 4, 1998, pp. 1407-1423. DOI:10.1109/18.681318.

[14]. Goldberger A. L., Amaral L. A. N., Glass L., Hausdorff J. M., Ivanov P. C., Mark R. G., Mietus J. E., Moody G. B., Peng C. K., Stanley H. E. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation, vol. 101, issue 23, 2000, pp. e215-e220. DOI: 10.1161/01.CIR.101.23.e21.

[15]. Moody G. A new method for detecting atrial fibrillation using RR intervals. Proc. Comput. Cardiol., vol. 10, 1983, pp. 227-230.

[16]. Chicco D., Jurman G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Mining, vol. 16, issue 1, 2023, p. 1. DOI:10.1186/s13040-023-00322-4.

[17]. Efron B. Bootstrap methods: another look at the jackknife. Breakthroughs in statistics: Methodology and distribution. In Breakthroughs in Statistics. Springer Series in Statistics. New York, Springer, 1992, pp. 569-593. DOI:10.1007/978-1-4612-4380-9_41.

[18]. Benjamini Y., Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), vol. 57, issue 1, 1995, pp. 289-300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

[19]. Andersen R. S., Poulsen E. S., Puthusserypady S. A novel approach for automatic detection of Atrial Fibrillation based on Inter Beat Intervals and Support Vector Machine. In 2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC). New York, IEEE, 2017. pp. 2039-2042. DOI:10.1109/EMBC.2017.8037253.

[20]. Xia Y., Wulan N., Wang K., Zhang H. Detecting atrial fibrillation by deep convolutional neural networks. Computers in biology and medicine, vol. 93, 2018, pp. 84-92. DOI: 10.1016/j.compbiomed.2017.12.007.

[21]. Wang Y., Yao Q., Kwok J. T., Ni L. M. Generalizing from a few examples: A survey on few-shot learning. ACM computing surveys, vol. 53, issue 3, 2020, pp. 1-34. DOI:10.1145/3386252.

[22]. Ng Y., Liao M. T., Chen T. L., Lee C. K., Chou C. Y., Wang W. Few-shot transfer learning for personalized atrial fibrillation detection using patient-based siamese network with single-lead ECG records. Artificial Intelligence in Medicine, vol. 144, 2023, p. 102644. DOI: 10.1016/j.artmed.2023.102644.

[23]. Abdelfattah M. S., Hagiescu A., Singh D. Gzip on a chip: High performance lossless data compression on fpgas using opencl. In Proceedings of the international workshop on openCL 2013 & 2014. New York, ACM, 2014. pp. 1-9. DOI: 10.1145/2664666.2664670.

[24]. Lee K.-S., Park H.-J., Kim J.E., Kim H.J., Chon S., Kim S., Jang J., Kim J.-K., Jang S., Gil Y., Ho S.S. Compressed deep learning to classify arrhythmia in an embedded wearable device. Sensors. vol. 22, issue 5, 2022, p. 1776. DOI: 10.3390/s22051776.

[25]. Yang Z., Yu Y., You C., Steinhardt J., Ma Y. Rethinking bias-variance trade-off for generalization of neural networks. In Proceedings of the 37th International Conference on Machine Learning. Brookline, JMLR, 2020. pp. 10767-10777. URL:https://proceedings.mlr.press/v119/yang20j.html.

[26]. Abrosimova M., Rebak A., Novikov R., Markov N. Classification with PPMd Compression in Few-Shot Learning: The Case of Eye-Tracking Dyslexia Detection. In 2024 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT). New York, IEEE, 2024. pp. 204-207. DOI:10.1109/USBEREIT61901.2024.10583975.

Информация об авторах / Information about authors

Никита Сергеевич МАРКОВ – младший научный сотрудник лаборатории трансляционной медицины и биоинформатики Института иммунологии и физиологии УрО РАН, ассистент кафедры вычислительной математики и компьютерных наук Уральского федерального университета. Сфера научных интересов: машинное обучение в физиологии и медицине, машинное обучение с малым количеством проб, обработка биомедицинских сигналов, высокопроизводительные вычисления, математическое моделирование в биофизике.

Nikita Sergeevich MARKOV – junior researcher at the Laboratory of Translational Medicine and Bioinformatics of the Institute of Immunology and Physiology UrB RAS, assistant at the Department of Computational Mathematics and Computer Science of the Ural Federal University. Research interests: machine learning in physiology and medicine, few-shot learning, biomedical signal processing, high-performance computations, mathematical modeling in biophysics.