Expert-Enhanced Machine Learning for Cardiac Arrhythmia Classification

Sebastian Sager[a,e,*], Felix Bernhardt[a], Florian Kehrle[b,e], Maximilian Merkert[a], Andreas Potschka[d], Benjamin Meder[b,e], Hugo Katus[b,c,e], Eberhard Scholz[b,e]

[a] *Otto-von-Guericke University, Department of Mathematics, Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany*
[b] *University Hospital Heidelberg, Department of Internal Medicine III, Im Neuenheimer Feld 410, 69120 Heidelberg, Germany*
[c] *German Centre for Cardiovascular Research, Im Neuenheimer Feld 410, 69120 Heidelberg, Germany*
[d] *Clausthal University of Technology, Institute of Mathematics, Erzstraße 1, 38678 Clausthal-Zellerfeld, Germany*
[e] *Informatics for Life, Im Neuenheimer Feld 669, 69120 Heidelberg, Germany*

Abstract

We propose a new method for the classification task of distinguishing atrial fibrillation (AFib) from regular atrial tachycardias including atrial flutter (AFlu) on the basis of a surface electrocardiogram (ECG). Although recently many approaches for an automatic classification of cardiac arrhythmia were proposed, to our knowledge none of them can distinguish between these two. We discuss reasons why deep learning might not yield satisfactory results for this task.

We generate new and clinically interpretable features using mathematical optimization for subsequent use within a machine learning (ML) model. These features are generated from the same input data and by solving an additional regression problem with complicated combinatorial substructures. The resulting model can thus also be seen as a completely novel ML model that incorporates expert knowledge on the pathophysiology of AFlu. Our approach achieved an unprecedented accuracy of 82.84% and an excellent area under the ROC curve of 0.9. One additional advantage of our approach is the inherent interpretability of the classification results. Our features give insight into a possibly occurring multi-level atrioventricular blocking mechanism, which might improve treatment decisions beyond the classification itself. Our research ideally complements existing cardiac arrhythmia classification methods from the literature, which can provide a preclassification, but so far left the important case AFib↔AFlu open.

*Corresponding author

Email addresses: `sager@ovgu.de` (Sebastian Sager), `felixbernhardt1992@gmx.de` (Felix Bernhardt), `florian.kehrle@med.uni-heidelberg.de` (Florian Kehrle), `maximilian.merkert@ovgu.de` (Maximilian Merkert), `andreas.potschka@tu-clausthal.de` (Andreas Potschka), `Benjamin.Meder@med.uni-heidelberg.de` (Benjamin Meder), `Hugo.Katus@med.uni-heidelberg.de` (Hugo Katus), `Eberhard.Scholz@med.uni-heidelberg.de` (Eberhard Scholz)

Preprint submitted to Artificial Intelligence in Medicine *November 15, 2020*

Keywords: Cardiac arrhythmia, Mathematical optimization, Machine learning, Decision support, Combinatorial algorithms, Classification problem.
2010 MSC: 92C50, 92B25, 90C27, 90C90, 68T05

1. Introduction

1.1. Automatic Classification of Cardiac Arrhythmias

The recent success of ML algorithms to classify cardiac arrhythmias is impressive [1]. However, the authors of this survey state: *"A known limitation of current ML methods is that it is challenging to understand the rationale behind their results. The algorithms are not able to provide explanations for the pathophysiological basis of classification outcomes, as they are unable to reveal the functional dependencies between data inputs and classes."* We agree with this point of view. For example, it is usually not clear if the classification results [2, 3, 4, 5] were due to heart rate variability, to the particular shape of the electrocardiogram (ECG) curve (including low voltage flutter waves that correspond to atrial polarizations), or to a mixture of both. Wavelets have been used to extract features automatically [6], but this approach is so far limited to easy classification cases and does not directly provide physiologically interpretable features. Parameters like the atrial cycle length are usually not provided, although they might be relevant for treatment decisions [7].

Moreover, none of the surveyed studies addressed the difficult and important special case of AFib↔AFlu, i.e., atrial fibrillation (AFib) versus regular atrial arrhythmias including atrial flutter and focal atrial tachycardias with irregular ventricular response (which we summarize shortly as AFlu in the following). Either it is completely omitted as in [6], which focuses on the classification classes normal beat, left bundle branch block beat, right bundle branch block beat, atrial premature beat, paced beat, and premature ventricular contraction. Or both physiological cases are lumped together in deep learning, *"The atrial fibrillation class combined atrial fibrillation and atrial flutter"* [3], and in algorithms based on heart rate variability for smartwatches [8]. Also studies that explicitly address "detection of AFib" in the title [9, 10, 11] can only detect the lumped class of irregular ventricular response which may either be due to AFib or to AFlu. The reason for this is that the special case AFib↔AFlu is difficult. The typically available data, a surface ECG or a time series of heart beats, look very similar in both cases to most laymen, physicians, and computerized algorithms alike. High misdiagnosis rates and possible causes have been reported [12, 13, 14]. This is concerning, as different treatments (often antiarrhythmics in AFib versus a highly successful ablation therapy in AFlu) are implied by the diagnosis [15] and atypical forms of AFlu are becoming increasingly important in clinical practice as a complication of left atrial ablation procedures [16]. See [17] for a more detailed discussion. The poor quality of expert opinion due to the difficult discrimination poses also a challenge to automatized classification by supervised ML, which often uses it for labeling training samples [3, 4, 5]. We used an expert analysis based on intracardiac measurements as gold standard, which is only available with invasive procedures.

Interestingly, the case AFib↔AFlu seems to be also difficult for deep learning approaches. As stated before, the differentiation between AFib and AFlu has been avoided in [3], where a deep convolutional net with 34 layers was trained using 91232 singe-lead ECGs. Also our results show poor performance of neural-network-based approaches. We conjecture that this might be due to the non-continuous nature of the underlying process which contrasts to the approximation properties of deep neural networks.

1.2. Complementing Previous Work in Automatic Arrhythmia Classification

Figure 1 visualizes our workflow. Deep learning (DL) can robustly distinguish samples that are either AFib or AFlu from sinus rhythm and 12 cardiac arrhythmias [3] with high accuracy. Other studies achieved similar results [6, 9, 10, 11]. As a reliable preclassification (Phase 0) can thus be achieved automatically (or manually), we focus here on Phase 1 (generation of physiologically interpretable features) and Phase 2 (using them for AFib↔AFlu classification).

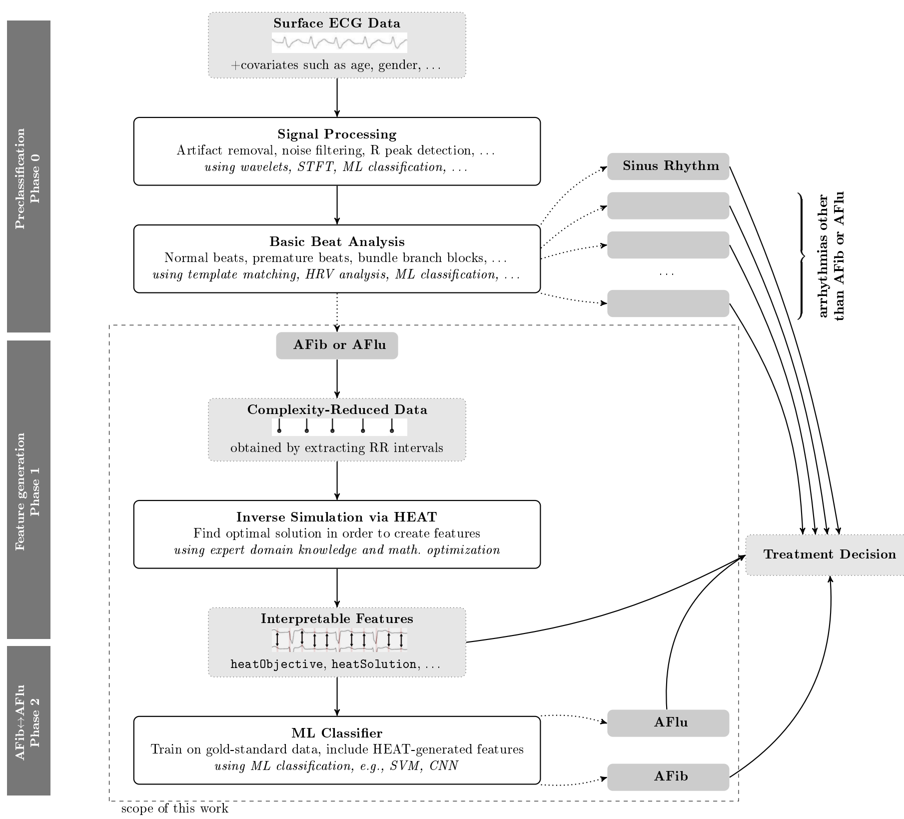

Figure 1: Visualization of our workflow from surface ECG to decision support for treatment.

Thus, in the following we are going to assume that it has already been verified that exclusively either AFib or AFlu is present, which is also true for our gold-standard data set (expert classification of intracardiac measurements which are only available after invasive procedures).

We propose to extend and complement the mentioned approaches with generated features that are based on a pathophysiological rationale allowing to also classify AFib$\leftrightarrow$AFlu. Thus our approach is not an alternative to previous work of automatic classification, but can be seen as an extension. In previous work, neural networks were trained with genetic algorithms [6] or with tailored stochastic gradient methods [3]. Our approach differs as it uses optimization in two different phases. In Phase 1, features are generated solving mixed-integer optimization problems. In Phase 2, an automatic classification is calculated using optimization. This approach is very modular and in general any classification algorithm can be applied in Phase 2.

1.3. Feature Generation and Hybrid Modeling

Feature construction has a long history, with early work dating back to the 1960s [18], and a plethora of feature generations methods, such as polynomial [19], by discretization [20, 21], by normalization [22], or grouping operations involving min, max, averaging, etc. The current state of the art in feature construction, however, suffers from three main drawbacks: exponential explosion of the feature space, difficulty to embed domain knowledge, and loss of interpretability. While the first drawback can be mitigated by feature selection methods, which can themselves be based on machine learning technology [23], the difficulty to embed domain knowledge and to interpret the automatically generated and selected features still remains. Our proposed feature generation does not suffer from any of the three drawbacks. Because it is based on the idea to embed domain knowledge (distilled into a mathematical optimization model), the generated features provide insightful interpretation to medical practitioners (but probably not to laymen), and exponential explosion of the feature set is not an issue because only few additional real-valued features need to be added.

As our feature generation procedure uses only the input data (RR interval times) and is also based on optimization, the whole procedure can also be seen as a completely novel machine learning model, with a nested hybrid structure. On the outer level it contains a classical ML part such as a Support Vector Machine (SVM), and in the inner part an inverse simulation domain knowledge model. The optimization on the outer level interacts with the results of the optimization on the inner level.

Combining machine learning models with domain knowledge is an active and promising field of research. A survey how first principle models can be combined in different ways with generic machine learning models is given in [24] in the context of process engineering systems. One promising way is to replace uncertain parts in differential equations with neural nets using the concept of universal differential equations [25]. Machine learning can also be applied to make the solution of differential equations more efficient [26]. The alternative is to develop and use physics-informed or biology-informed machine learning approaches [27, 28, 29, 30, 31]. The general idea is to design machine learning models such that important physical properties like conservation laws are automatically fulfilled. This promising line of research is often linked to the simulation of complex flows. A physics-informed neural network was applied to real noisy clinical data in [32]. Here, arterial pressure was predicted from MRI data of blood velocity and wall displacement. A common result of these studies is that by combining physics-based and machine learning models, it is often possible not only to improve the performance of the purely black-box machine learning models, but also to make them more transparent and interpretable.

The mathematical model that we develop and apply in this paper can be seen as a simplification of first-principle models for electrical conductivity in the heart, such as the Hodgkin-Huxley equations [33]. In this sense, our approach can also be interpreted as a biology-informed machine learning approach. See [34] for a survey of systems biology models and important properties.

1.4. Summary of Our Approach

The most important building block in Phase 1 is the inclusion of medical expert knowledge. It was unclear for a long time which role the atrioventricular (AV) node played in the transfer of fast but regular activations of the atrial chambers into irregular activations of the ventricular chambers. Or as Douglas P. Zipes stated in 2000, the AV node is still *"a riddle wrapped in a mystery inside an enigma"* [35]. Key to solving this riddle is the idea of a multi-level AV block (MAVB) [36, 37, 38, 39, 40]. The tedious procedure of manually adjusting possible MAVB combinations has been automatized with large success in the algorithm HEAT (Heidelberg Electrocardiogram Analysis Tool, [17]). The underlying hypothesis is that fast but regular activations of the atrial chambers result in irregular responses of the ventricles because of a (multilevel) succession of simple blocks of *Type I* or *II*. We considered atrial cycle length, blocktype, and a vector of blocktype-specific internal offset counters and conduction constants as optimization variables. For different values of these variables, forward simulation of ventricular responses (RR interval lengths) is possible, which can be compared to given RR measurements. A penalization of the difference in an appropriate metric gives a suitable objective function. In an inverse simulation, HEAT can calculate optimal solutions resulting in the smallest deviations for all training samples. The combination of mathematical model and optimization algorithm could also be seen as an interpretable expert system. The basic idea of using a mathematical model and inverse simulation for AFib$\leftrightarrow$AFlu classification has been published before [17]. We here report a significantly matured approach with a larger data set ($4\times$) which allowed a systematic cross-validation, an improved mathematical model of MAVB with a better pathophysiological interpretation, a computational speed up ($5000\times$), and an increased accuracy. Most importantly, for the first time we use HEAT for multi-dimensional ML feature generation and show the advantages of using clinical domain knowledge. The general approach to use domain knowledge plus combinatorial optimization for feature generation might overcome intrinsic approximation limits of deep learning for non-smooth systems, as they often occur in medicine and biology, e.g., [41, 42, 43, 44].

1.5. Organization of this Paper

The paper is organized as follows. In Section 2 we describe our machine learning approach and data. In particular, we explain a mathematical model that is used as domain knowledge to describe AFlu and derived features. In Section 3 we present numerical results that show that the proposed approach reaches an unprecedented accuracy, while a direct use of neural networks perform poorly on the data. In Section 4 we discuss these results in several directions: approximation properties of machine learning as a possible explanation, accuracy and impact, interpretability, and transfer to other clinical domains. Concluding remarks are given in Section 5.

2. Methods

2.1. Multilevel Atrioventricular Block (MAVB)

We developed a mathematical model for MAVB based on the following rationale. In physiology, *refractoriness* specifies the time period in which a cell is incapable of repeating a certain action. Applied to any component in the cardiac conduction system, one distinguishes the *absolute refractory period* (ARP) describing the duration in which a cell can not be stimulated under any circumstances and the *relative refractory period* (RRP) describing the duration in which the tissue can be stimulated under certain conditions, but may react with a modified conduction [45]. Depending on incoming signal and RRP, a block ratio of $n+1:n$ can occur, where $n+1$ is the number of incoming, and n the number of conducted signals. Due to changes in cell fatigue or in the frequency of the incoming signals, this ratio may vary, even on short time horizons. For larger values of n the conduction times may change as well.

Motivated by the physiology of the AV node, we considered it as a series of cell compounds in which a signal may potentially be blocked. Hence, the outgoing signal of block level I becomes the incoming signal of block level II, and so on, see Figure 2. Classifying atrial flutter with irregular ventricular response (AFlu, left) versus atrial fibrillation (AFib, right) based on the surface electrocardiogram (ECG, bottom) is difficult for experts and algorithms. If intracardiac measurements were available (after invasive procedures, like in our data set), the classification would be easier (regular versus irregular, top row of the figure), allowing to use it as a gold standard for training of machine learning models and for a-posteriori analysis. The input data of the feature generation, the measured ventricular (V) signals (`rawRR`), were extracted from the surface ECG (bottom of figure). For both samples a two-level atrioventricular (AV) block was calculated such that the model parameter Δa, the cycle length in the atrial chambers (A), is regular and the forward simulation in V is close to `rawRR`. We hypothesized that a small deviation (left) can be interpreted as a high likelihood for regular behavior (AFlu), and a large deviation (right) for

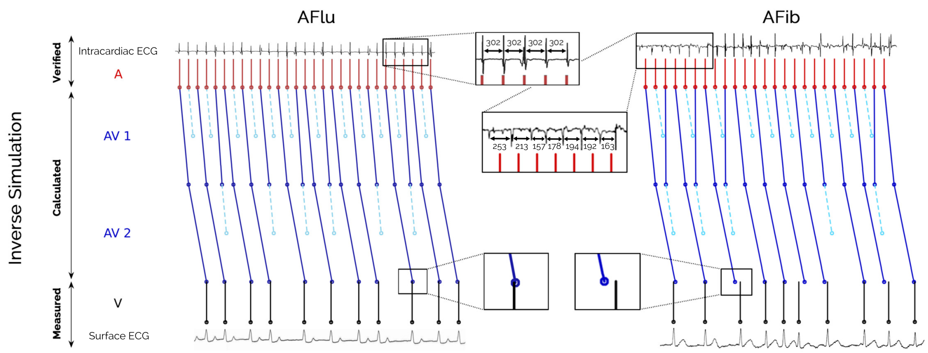

Figure 2: Visualization of our inverse simulation approach applied to two samples.

chaotic behavior which can not be explained well by the model (AFib), compare bottom zooms in the figure, cf. [17]. It can be visually confirmed that for AFlu the calculated Δa corresponds well to the intracardiac measurements.

This theoretical concept allows to combine different blocking ratios $n+1:n$ (possibly varying and with linearly changing conduction times due to RRP, denoted as *Type I*) on an unlimited number of levels. However, it makes sense to limit the number of possible combinations to avoid overfitting, to reduce computational time, and to stay close to clinical observations. We restricted our MAVB model to the five combinations shown in Figure 3 with a maximum of three block levels, consistent with cases described in the current literature.

The resulting mathematical model comprises most different classical and advanced block types, in particular typical Type I block [46, 47, 48], atypical Type I block [47, 49], the special cases of 2:1 and 3:2 Type I blocks, Type II block [50, 51, 52, 53], advanced second-degree AV Block [54, 55], and MAVB [36, 37, 38, 39, 40]. Preferable in the sense of Occam's razor, this unified model also allows an efficient calculation of the most likely block for given RR data.

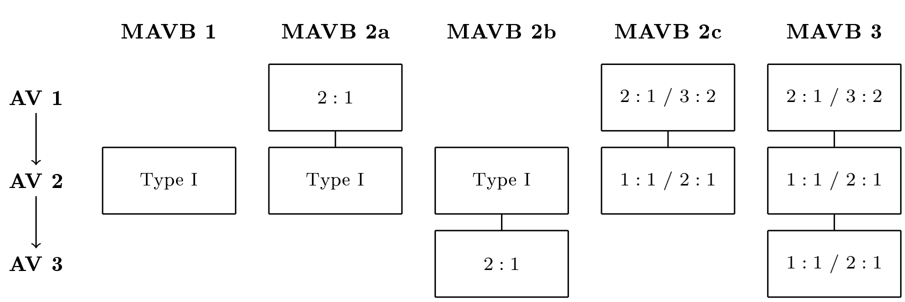

Figure 3: The blocktype *bt* can be chosen as one of the five depicted combinations of up to three multilevel atrioventricular block (MAVB) levels.

2.2. HEAT

For the inverse simulation optimization problem we considered optimization variables $x = (\Delta a,\, bt, oc)$, where Δa is the atrial cycle length, bt the blocktype, and oc a vector of auxiliary variables representing blocktype-specific internal offset counters and conduction constants. The objective function is denoted by F_i where $F_i(x)$ measures the deviation of the result of the forward simulation based on x from the actual RR data sample i in the Euclidean norm.

With the help of the software package HEAT we calculated for all training samples i optimal solutions x_i^*, i.e., particular values for Δa_i^*, bt_i^*, and oc_i^* which resulted in the smallest objective function value

$$F_i(x_i^*) = \min_{x \in \mathcal{X}} F_i(x).$$

Here $\mathcal{X}$ denotes the feasible set for $(\Delta a, bt, oc)$ with lower and upper bounds for $(\Delta a,\, oc)$ and five possible blocktypes that comprise most clinically observed types of MAVB, compare Figure 3. The bounds on the atrial cycle length Δa were determined based on physiological observations [45] (between 175ms and 400ms) and dependent on the blocktype bt and the input RR data. The algorithm is based on an intelligent enumeration (comparable to Dynamic Programming or Branch&Bound) of all possible solutions, assuming a time grid of 1ms for Δa and oc. The proprietary software and the data set `heatDS` are available for academic studies on reasonable request.

2.3. Features and Feature Sets

As features, we investigated the time series of raw input RR interval times (`RR`), together with the derived scalar features heart rate variability (`RRvar`) and average heart rate (`RRmean`); the HEAT optimal objective function value $F(x^*)$ (`HEATobj`); and the HEAT optimal solution (variable assignments) $x^* = (\Delta a^*,\, bt^*, oc^*)$ (`HEATsol`).

To further increase accuracy and stability, we applied a moving horizon strategy to generate additional features as follows. From the $n_{\mathrm{RR}} = 22$ time intervals, we considered only $n_{\mathrm{sub}} \in \mathcal{I} := \{10, \ldots, n_{\mathrm{RR}}\}$ on windows $[1, 2, \ldots, n_{\mathrm{sub}}]$ until $[n_{\mathrm{RR}} - n_{\mathrm{sub}} + 1, 2, \ldots, n_{\mathrm{RR}}]$. This results in additional solutions $F_{i,n_{\mathrm{sub}}}(x^*_{i,n_{\mathrm{sub}}})$ for $i \in \mathcal{I}$. To investigate the robustness of solutions, we also evaluated $F_{i,j}(x^*_{i,k})$ for $j, k \in \mathcal{I}$, i.e., how well do optimal solutions of time window j perform on time window k. We thus computed the features `HEATobj` and `HEATsol` for each subwindow of RR intervals. The moving horizon approach also enabled us to use a comparison of the HEAT simulation based on one time window with the raw RR intervals of a different one, as described above ("how well performed optimal solutions of time window j on time window i"?) (`HEATfit`). We refer to the resulting time series of $n_{\mathrm{RR}} - n_{\mathrm{sub}} + 1$ entries `HEATobj`, `HEATsol`, and `HEATfit` as `HEATseries`, to the generically derived features mean and standard deviation as `HEATseriesAvg`. Finally, we also considered patient age (`age`). Table 1 summarizes the sets of features and resulting dimensions.

Feature Set	included **Features** **ML Model**	**# Pars**	**# Scalings**	**# Hyp**
`rawRR`	$=\{$`RR`$\}$			
	CNN	287–487	0	2
	SVM N-Gram	101–485	200–968	4
`heatObjective`	$=\{$`HEATobj`$\}$			
	SVM	2	2	4
`heatSolution`	$=\{$`HEATobj`, `HEATsol`, `RRvar`, `RRmean`$\}$			
	SVM	10	18	4
`heatSerAvg`	$=\{$`HEATseriesAvg`$\}$			
	SVM	21	40	4
`heatSerAvgAge`	$=\{$`HEATseriesAvg`, `age`$\}$			
	SVM	23	44	4
`heatSeries`	$=\{$`HEATseries`$\}$			
	SVM N-Gram	91–1691	180–3380	4

Table 1: Number of optimization parameters (Pars), scaling factors, and hyperparameters (Hyp) for the different feature sets and ML models.

2.4. ML Models

We used two classes of standard ML classification models, namely support vector machines (SVM) and convolutional neural networks (CNN).

Since a SVM does not incorporate the temporal connection between sequential data, we first computed general features based on subsequences (N-Grams) of the underlying data. These general features are the mean and the standard deviation of a given subsequence. For the mean, any subsequence with length ≥ 1 and $\leq n_{\mathrm{RR}}$ was considered. The standard deviation was only computed on subsequences of length ≥ 2. The hyperparameter n_{sub} limits the length of the time series before computing the features. Before being used for training, each feature was standardized to zero mean and unit standard deviation. The necessary parameters for this transformation were computed on the training set and also used for the model evaluation. Based on these features, we implemented a SVM model in scikit-learn based on the LIBSVM library [56]. The underlying model is described in [57]. The kernel type (radial basis functions or polynomial) with a penalty parameter C and a kernel coefficient γ (3 values each) and the length of analyzed subsequences $n_{\mathrm{sub}} \in \{10, \ldots, 22\}$ were tuned as hyperparameters using grid search cross-validation.

We used a CNN architecture consisting of 2 convolutional blocks followed by 1 fully connected layer with rectified linear unit (ReLU) activation functions and 1 final fully connected layer with a sigmoid activation function and output dimension 1. Each of the convolutional blocks consisted of 2 convolutional layers with ReLU activation functions and 5 filters of width 2 followed by a max pooling and a dropout layer. The dropout rate (10%, 20%, 30%) and n_{sub} were tuned as hyperparameters during training using grid search cross-validation.

Other objective functions and architectures were evaluated manually in a preliminary phase, but not further considered as they gave no additional insight.

Table 1 shows the number of optimization parameters, of scaling factors, and of hyperparameters for the different approaches. The number of optimized parameters may depend on the hyperparameter n_{sub} (the length of analyzed subsequences), therefore also ranges are provided. To avoid overfitting, each approach was evaluated on `heatDS` using repeated, stratified 10-fold cross validation to estimate performance on new data.

2.5. Data

Our data set `heatDS` is a superset of the one used in a previous study [17], which contains details concerning the data obtained from patients exhibiting AFib or AFlu with irregular ventricular response during invasive electrophysiological testing or catheter ablation. The retrospective data was extended to the period between 2011 and 2018 and a total of 159 patients.

For all 159 patients the classification AFib↔AFlu was performed using electrical signals measured at the atrial electrodes by an expert in the field of cardiac electrophysiology. For AFib, we found that all examples exhibit highly irregular intervals of atrial activation (qualitative assessment) in combination with a short mean atrial cycle length (Δa) of 182 ms. These data correspond well with the threshold of 200 ms that is referred to in the European guideline for the management of AFib [58]. In contrast, intracardiac recordings taken from patients with AFlu exhibited highly regular intervals ($\Delta a \approx 240$ ms). In many cases, the correct rhythm diagnosis could be verified by evaluating the reaction of the arrhythmia to catheter ablation. Among the group of AFlu cases, further quantitative assessment revealed a Δa variation below 5 ms.

Our hypothesis was that the dynamics of ventricular activations in short time periods contain enough information for a successful discrimination. Therefore we reduced the data complexity by extracting the time interval durations of 22 RR intervals from the surface ECG using built-in calipers, to a precision of 1 ms. Segments containing premature ventricular beats were excluded.

In summary, we collected 380 examples which were diagnosed either AFlu ($n = 190$) or AFib ($n = 190$). We used either two or three disjoint examples per patient to increase the overall data size. We stored the time series of 22 values corresponding to RR intervals, the patient's age, and the correct label AFib/AFlu for training and validation purposes. All other ECG data (including the intracardiac measurements) were not considered further, except for exemplary a-posteriori illustration. The study was approved by the ethics committee of the University of Heidelberg and conforms to the standards defined in the Helsinki Declaration.

In [59], we validated a previous version of our algorithm also against other, smaller data sets from the literature which focused on AFib↔AFlu discrimination. Unfortunately, there are no larger data sets available that can be used as an extended benchmark. Usually, these either don't differentiate between AFib and AFlu in specific or they do not classify supraventricular tachycardias at all, like the American Heart Association ECG Database for example [60]. E.g., all of the data in the studies [8, 9, 10, 11] is of no use for us, as it is unlabeled with respect to AFib↔AFlu.

2.6. Implementation Setting

All results were computed on a server running Ubuntu 16.04.4. The system had access to 1 TB RAM, an Intel(R) Xeon(R) CPU E5-2699A v4 at 2.40GHz with 88 cores, and two NVIDIA(R) Quadro(R) p5000. The ML models were implemented using Python 3.5.2 and scikit-learn 0.20.3. The CNNs were based on tensorflow 1.8.0 and trained using the Adam optimizer [61] with default parameters. The computational times were roughly 20 milliseconds per HEAT call (times 380 samples times number of considered subproblems per sample), 30 minutes for training SVM, and 3 days for training CNN.

3. Results

3.1. Accuracies for Different Feature Sets and ML Models

We show the mean averages (averaged sensitivity and specificity) and areas under receiver operating characteristic curves in Table 2. The results were obtained after repeated, stratified 10-fold cross validation for different feature sets and ML models as described in Sections 2.3 and 2.4.

When directly applied to the input data of at most 22 RR interval times (`rawRR`), standard ML approaches achieved approximately 60%. The average accuracy increased to 77.58%, when $F_i(x_i^*)$ was used as the only feature (generated a priori from `rawRR`). A higher-dimensional classification, which also took x_i^* and several HEAT solutions from a moving horizon strategy into account, increased the average accuracies to 79.37% and 82.84%, respectively. Using the best approach, we achieved a sensitivity of 87.21% and a specificity of 78.47%. An exemplary distribution of features is shown in Figure 6.

For an implementation of a convolutional neural network (CNN) the poor performance of direct application to `rawRR` was also reflected by high standard deviations. The number of ML parameters was two orders of magnitude larger

Feature Set	ML Model	Accuracy	ROC Area
`rawRR`	CNN	57.26% ± 6.47%	0.60 ± 0.08
	SVM N-Gram	62.03% ± 5.25%	0.66 ± 0.07
`heatObjective`	SVM	77.58% ± 4.15%	0.85 ± 0.05
`heatSolution`	SVM	79.37% ± 4.55%	0.87 ± 0.03
`heatSerAvg`	SVM	82.18% ± 4.48%	0.89 ± 0.03
`heatSerAvgAge`	SVM	82.47% ± 3.26%	0.90 ± 0.03
`heatSeries`	SVM N-Gram	82.84% ± 4.31%	0.90 ± 0.04

Table 2: Average accuracies and Areas under Receiver Operating Characteristic (ROC) curve with standard deviations for the different approaches.

than for SVM, although only few layers were chosen due to the small size of the training set and compared to DL approaches to cardiac arrhythmia classification [3]. The SVM results were quite stable. E.g., no significant differences occurred for different kernel types. The approach to preprocess `rawRR` using medical expert knowledge (HEAT) can thus also be seen as an approach to increase sensitivity without overfitting the ML model.

3.2. *Interpretability*

Whereas we observed that the calculated objective function values $F_i(x_i^*)$ were the most decisive feature for classification, the features associated with x_i^* are interesting from a clinical interpretation point of view. Figure 4 shows how the knowledge of the atrial cycle length Δa^* might be helpful for an a-posteriori identification of flutter waves for AFlu in a surface ECG. The figure shows observed and simulated data, as in Figure 2 left, but for different input data from the same patient. The actual atrial cycle length is only available with invasive procedures and is difficult to identify from investigating the surface electrocardiogram (ECG, rightmost zoom), where almost no atrial activation is recognizable. The intracardiac measurements are shown for illustrative purposes and coincide with the value Δa proposed by HEAT (leftmost zoom). When no intracardiac measurements are available, this value Δa could be of help for the physician, e.g., when carefully reanalyzing the ECG. An overlay of Δa makes the task to spot atrial activations in the surface ECG easier (middle zoom).

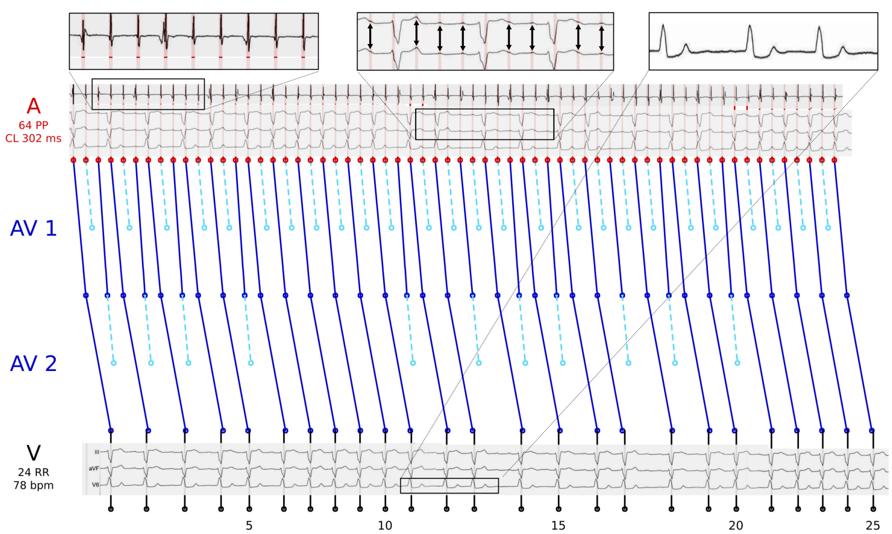

Figure 4: Exemplary illustration of how the feature *atrial cycle length* derived from a HEAT solution can be a posteriori pathophysiologically interpreted and used.

Figure 5 again shows observed and simulated data, but for different input data. Here, a three-level atrioventricular (AV) block with a varying 2:1 / 3:2

level followed by two levels with a varying 1:1 / 2:1 conduction was calculated (MAVB 3 in Figure 3). Again, the intracardiac measurements are shown for illustrative purposes (top). The close match to the calculated atrial cycle length Δa highlights the plausibility of the complex blocking mechanism. The optimal blocktypes bt^*, compare Figures 4 and 5 with two and three levels with varying blockings, respectively, give insight into the pathophysiology of the AV node and might be useful for choosing a good treatment.

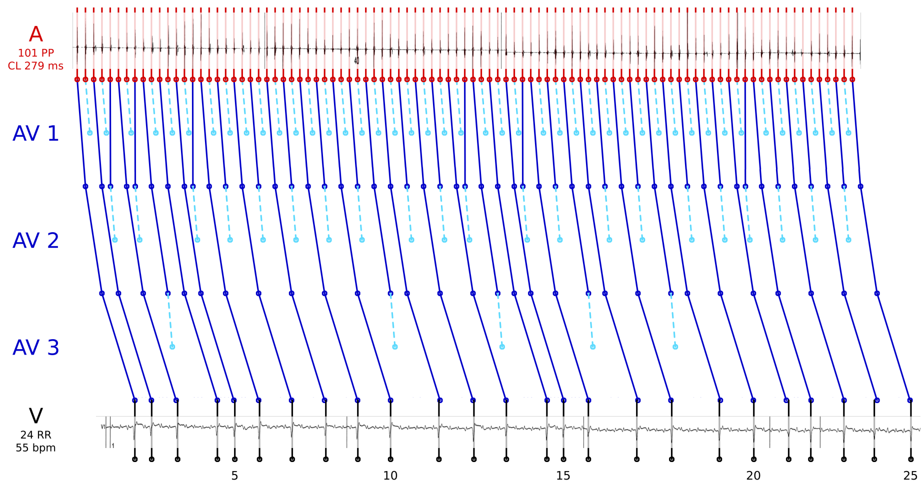

Figure 5: Exemplary illustration of how the feature *blocktype* derived from a HEAT solution can be a posteriori pathophysiologically interpreted and used.

The high accuracy of ML approaches that used HEAT-generated features indicates that our novel mathematical model is an appropriate description of the complex blocking mechanism for AFlu.

3.3. Moving Horizon Approach

The results in Table 2 seem to indicate that additional accuracy can be obtained by using the feature `HEATseries`. It consists of time series data generated from several calls to HEAT for input data obtained from a moving horizon approach. As explained above, $n_{\mathrm{sub}} \in \{10, \ldots, n_{\mathrm{RR}}\}$ was optimized as a hyper-parameter, with $n_{\mathrm{sub}} = 17$ giving the best results. The overall number of time intervals $n_{\mathrm{RR}} = 22$ was fixed. Therefore, the time series in `HEATseries` corresponded to entries for 6 different optimization problems $(1 \ldots 17$ to $6 \ldots 22)$.

An interesting and promising question is, if and how much the approach can be improved for larger values of n_{RR}. Unfortunately, the idea to use several optimization results in one feature set came up later in the project, when data from many patients was already collected, with small numbers of RR intervals. Considering the collected number of RR intervals for the 159 patients, the average number is 51 with a range from 22 to 111. This made a rigorous cross-validated comparison of larger values of n_{RR} difficult, as our data base was simply not large enough. However, a study showed large potential: the accuracy rose from 82.94% to 92.50% for long time horizons of $n_{\mathrm{RR}} = 90$ intervals. However, this result needs to be cross-validated on larger data sets.

4. Discussion

4.1. Impact, Accuracy, and Applicability

Being able to classify AFib↔AFlu is clinically relevant. There are a variety of treatments (antiarrhythmics, different kinds of ablations and ablation systems) with different side effects and chances for curing the patient. A correct classification is imperative to choose the best treatment [15]. Therefore a usage of the proposed approach for clinical decision support might be of great help, especially when considering the excellent classification accuracy and interpretability of calculated features on the one hand, and the difficulty of the classification task for unexperienced clinicians on the other hand.

All ML approaches that were applied directly to the input data (`rawRR`) resulted in average accuracies of approximately 60%. These low accuracies were not surprising, as AFib↔AFlu is a difficult case even for experts [12, 13, 14] and was explicitly excluded in recent studies [3]. AFib may be overdiagnosed because of coarse fibrillatory waves which are reminiscent of AFlu [13, 62], the presence of artifacts, or premature atrial complexes [63]. AFlu may be overdiagnosed because the low-voltage flutter waves that indicate AFib can be hardly discernible in the surface ECG, compare Figures 2 and 4, or because a pseudoregularization may occur [64], see also Section 4.5. The achieved accuracies are similar to previous results to analyze AFib↔AFlu, e.g., based on clustering of RR times or nodal recovery approaches [59]. Note that the N-Gram approach implicitly considers `RRvar`, `RRmean` and is thus a superset of features used in current smartwatch algorithms [8]. Hence, the low accuracy gives a hint why AFib↔AFlu can not currently be treated by them.

Using HEAT for an a-priori calculation of `heatObjective` was significantly more successful with an average accuracy of 77.58%, although the input data was identical (`rawRR`). Using `heatSolution` features resulted in an increased average accuracy of 82.84% (sensitivity 87.21%). Further improvements can be expected if settings of the HEAT algorithm (such as a lower bound on Δa or grid sizes) were optimized as hyperparameters, if underlying model assumptions were adapted after careful analysis of wrongly classified samples, once more training samples become available, and if covariates were considered. Age (`heatSerAvgAge`) did not seem to have a significant impact on accuracy, though.

Using ML with HEAT-generated features has the drawback that for every classification sample an optimal solution of the MAVB needs to be calculated. However, the additional runtime of 20 milliseconds should be acceptable in a clinical context and will be outweighed by several advantages.

First, the approach is applicable in clinical practice. We assumed that in a previous assessment the presence of either AFib or AFlu was verified. Seen from another angle, our approach is a reasonable complement to generic DL approaches for cardiac arrhythmias [3]. It can use the prior classification of AFib and AFlu into one cluster, and can classify AFib↔AFlu in a following step. HEAT can be run on a server. A secure client-server architecture has been implemented [59]. It allows communication with a smartphone app that generates `rawRR` data from ECG-derived pictures or beeps from a heart monitor. A similar procedure could be implemented for wearables and smartwatches.

Second, the dominance of the `HEATobj` feature and the availability of a distribution, compare Figure 6, allow calculation of a probability for the classification (the higher the value, the more likely AFib). Such a value would help clinicians to estimate the validity of the suggested diagnosis. From Figure 6 one observes the clear separation of atrial flutter (AFlu) and atrial fibrillation (AFib) with respect to `HEATobj`. The two model parameters in x^*, the atrial cycle length Δa and the blocktype bt do not allow a straightforward classification.

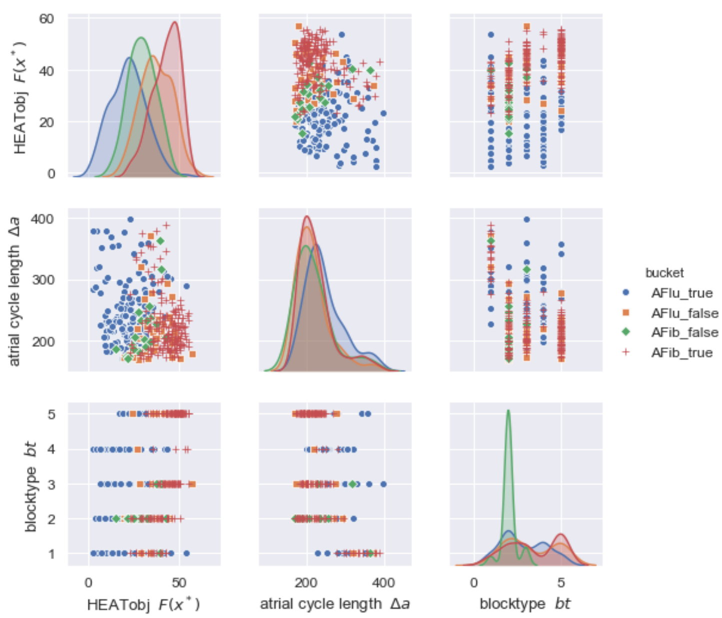

Figure 6: Representative pairwise plot of features obtained from a `heatSolution` SVM classification, compare Table 2.

Third and as discussed above, it results in a high accuracy. It is an open question whether a similar accuracy could be achieved with DL without the explicit modeling of expert knowledge. Probably yes, if the number of verified training samples, of hidden layers, and the computational resources were large enough, but even then the approach would lack interpretability.

4.2. Interpretability

Interpretability is the fourth and most important advantage of the proposed approach.

Led by text classification and image processing, machine learning has been conquering many areas of modern life and the sciences. Despite some disappointments [65], the combination of statistical modeling, optimization algorithms, increased computing power, open source initiatives, and availability of data has led to spectacular breakthroughs and an omnipresence of Artificial Intelligence in modern life and research, also in clinical information systems [66]. Yet, the unprecedented success of data-driven ML is accompanied by worries about acceptance, robustness of validation procedures, and interpretability of the results. These aspects are repeatedly named as main limitations of current AI systems demanding further research [67], in particular in healthcare applications [68, 69, 70]. Transparency and interpretability are explicit goals of national research programs. For instance, according to the National Artificial Intelligence Research and Development Strategic Plan of the US *"A key research challenge is increasing the 'explainability' or 'transparency' of AI. Many algorithms, including those based on deep learning (DL), are opaque to users, with few existing mechanisms for explaining their results. This is especially problematic for domains such as healthcare, where doctors need explanations to justify a particular diagnosis or a course of treatment."* [71]. Similar statements can be found in the German national AI strategy report [72].

We reduced the complexity of the data a priori by considering only time points of the clearly visible R waves (the beeps of a heart rate monitor) corresponding to ventricular activation. This makes the underlying data more assessable to humans. HEAT provides also `HEATsol`, i.e., the optimal solution $x^* = (\Delta a^*,\ bt^*, oc^*)$. These values can be interpreted by experts, and used for the treatment decision making. For example, the atrial cycle length Δa^* proposed by HEAT could be of help for the physician when carefully reanalyzing the ECG, compare Figure 4. Furthermore, the absolute cycle length could help identifying patients with typical atrial flutter ($\Delta a \sim 200$ ms) or predicting procedural success [7]. In addition, for AFlu *"a thorough understanding of electrophysiological properties and anatomical landmarks is essential in achieving a successful ablation outcome and in reducing complication rates"* [73]. Sometimes it is even claimed that *"the classic ECG-based diagnoses of tachycardias and AFib are of little importance today because treatment is based on the direct management of the trigger mechanism"* [74]. We believe that estimates of the atrial cycle length or the blocktype, compare Figures 4 and 5, could be a valuable asset to clinical decision making.

4.3. Impact of ML Architectures and Feature Selection on Accuracy

Table 2 shows the accuracies for different machine learning architectures. After reasonable effort to investigate different architectures none resulted in an accuracy significantly above 60% when directly working with `rawRR`. We think that this is mainly due to the comparatively small amount of data samples and the difficulty to tailor standard ML architectures to the specific time series character of RR intervals. When the features that were generated using domain knowledge were additionally considered, SVM outperformed our CNN architectures, see the discussion in the next subsection. We expect a different behavior if neural network architectures were used that explicitly address time series, such as recurrent networks.

A key ingredient in the proposed approach is the generation of features via domain knowledge. We solved an inverse optimization problem for the mathematical MAVB model introduced in Section 2.1. This generic approach seems preferable not only for the aforementioned reason of interpretability, but also because it makes the cumbersome tailoring of a generic neural network architecture for the specific classification task obsolete. The classification in the low-dimensional feature space can be efficiently and accurately done with SVMs.

The selection of features was straightforward, as there are only few model parameters that are calculated along with the objective function value. The latter alone was decisive and would already be enough for a high-accuracy 1-dimensional linear classifier (i.e., using a simple threshold value), compare the entry for `heatObjective` in Table 2. The additional features that we considered in `heatSolution` did increase accuracy additionally, although we see the main benefit of block type, atrial cycle length, and conduction constants in the physiological interpretability. Future work should focus on a consideration of sets of optimal solutions and solutions on moving time horizons. In this context the impact of `heatSolution` might increase further.

4.4. Approximation Properties of Machine Learning Approaches

It is well known that feed-forward neural networks are universal approximators of continuous functions, if either the number of neurons on one hidden layer [75] or the number of layers for a fixed number of neurons per layer [76] may grow. However, it is also well known that these beautiful theoretical results come at the price of a potentially large number of weights distributed over the hidden layers of the neural net. Adaptive activation functions seem to have better approximation properties [77], but the main difficulty of current architectures should be the same. To get an idea why CNNs do not seem to perform well on AFib↔AFlu, for deep nets with 34 layers as in [3] as well as in our prototypical implementation, we analyze Figure 7.

It shows the feature `HEATobj`, i.e., the optimal objective function value $F_i(x)$ provided by HEAT, for 801 different artificial input vectors x. As input, 17 RR intervals of an exemplary patient were chosen. 16 of them are kept fixed, while one particular interval length in the middle was varied with deviations of -400ms to $+400$ ms in steps of 1ms. The plot shows locally quadratic behavior, which is due to the quadratic objective function (Euclidean norm). The discontinuities are due to clipping of solutions that result in deviations of more than 150ms between signals. The main take-away from the plot is that the minimal objective function value as a function of the input consists of many piecewise quadratic segments. Estimating the number of ReLU-induced linear segments necessary to approximate this important feature for classification, one easily reaches large numbers: assume 20 linear segments, and use $n_{\text{sub}} = 17$ as an exponent. Of course the feature `HEATobj` is only an approximation of the real process, but the mathematical modeling based on physiological knowledge and the high accuracy indicate that the real MAVB will show a similar behavior. Given the additional

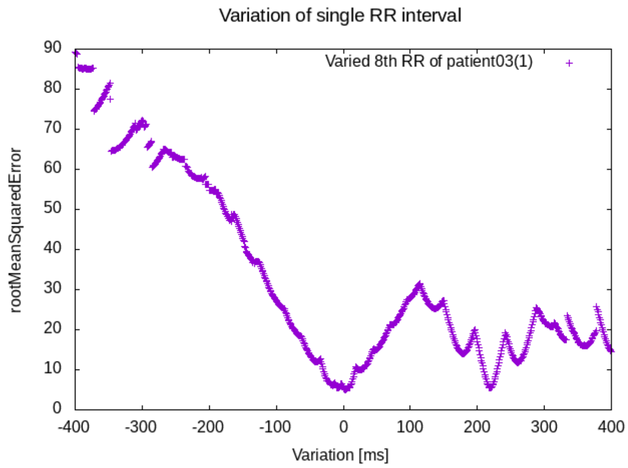

Figure 7: The objective function of our mathematical model fluctuates strongly with the input signal, a possible explanation why deep learning approaches yield poor approximations.

difficulty that for this difficult classification task only few labeled training data is available, we conjecture that it will be difficult to train CNNs with a reasonable classification accuracy without using domain knowledge.

4.5. Classification failures

While our novel approach resulted in excellent area under the curve values, there were still misclassification samples. Figure 8 shows an atrial fibrillation case with a very fast (160 beats per minute), but pseudoregular ventricular contraction, shown in the surface lead at the bottom. The atrial contraction on the other hand is totally chaotic as shown by intracardiac measurements displayed in the top. Due to this pseudoregularization, the best MAVB simulation matched the observed data quite well and led to a misclassification. It is well known that at very high frequencies of AFib a *pseudoregularization* can occur [64]. Here, the RR variability decreases with an increase in heart rate, which leads to an almost regular rhythm despite a totally chaotic atrial contraction. As a consequence, these AFib cases with high ventricular rates might be more likely to match a regular MAVB or even a $1:1$ conduction. In our approach pseudoregularizations result in relatively low objective function values which impair correct classification.

Just as for experts, the presence of artifacts or premature atrial complexes [63] might also lead to a misclassification. It is an open question how to extend the mathematical model in Section 2.1 such that pseudoregularization can be detected automatically and the overall specificity increases without impairing the sensitivity. Using the feature *atrial cycle length* in a more elaborate way or additionally classifying the flutter waves might be helpful in this context.

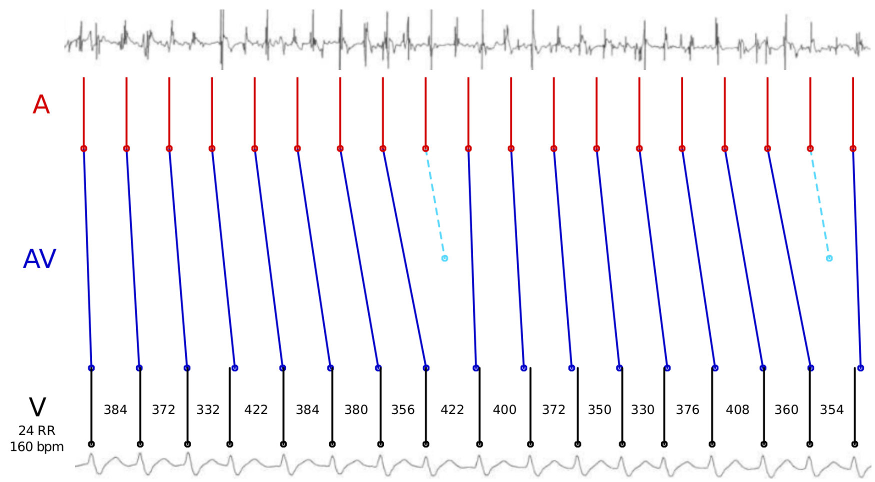

Figure 8: Example of a misclassification.

An intrinsic limitation for the classification accuracy using our approach arises from false positives, i.e., cases of AFib, that "by chance" are very close to multi-level blocks. The mathematical question of how dense random `rawRR` instances are in the space of all MAVB solutions is open.

4.6. Generalization to other cases of clinical decision support

Our proposed approach can be generalized as *"enhance ML approaches by features based on understandable and interpretable mathematical models of clinical expert knowledge that exhibit complex dynamic behavior"*. Personalizing these mathematical models results in model parameters that can be used for classification, prediction and dynamic stratification, but also interpreted by clinicians. Diagnosis of other cardiac arrhythmias could be done in a similar way as above. But also for diseases like acute leukemias [78, 79] or polycythemia vera [80] there are mathematical models which have been validated with measurement data, and which contain estimated personalized model parameters like stem cell proliferation rates. Such hidden parameters can usually not be observed directly and could be very useful for clinical decision making [81].

We believe that it is better to use interpretable models than to explain black box models [82]. An integration of interpretable expert systems written as optimization models with today's powerful ML approaches might result in better healthcare with interpretable results.

5. Conclusions

We proposed a method for the difficult classification task AFib↔AFlu that combines expert models and machine learning. On our test set of gold standard, our approach was highly successful and reached a classification accuracy of 82.84% and area under the ROC curve of 0.9. In contrast, for short RR time series and comparably few labeled training samples, we could not achieve such an accuracy with a purely data-driven ML model.

Our work ideally complements deep-learning-based methods, which can provide a preclassification, but can not further distinguish between AFib and AFlu. However, this distinction is highly relevant from a clinical perspective. The classification itself together with corresponding features calculated by HEAT may be interpreted by medical experts and utilized for the treatment decision. As runtimes of the algorithm are low enough for real-time requirements, it appears to be applicable as a decision-support tool for clinical practice.

An open question is how to further reduce failure cases due to so-called pseudoregularization as discussed in Subsection 4.5.

Finally, we proposed to create features from optimal solutions of domain-knowledge models and to search for unknown patterns in a lower-dimensional feature space. We think that this general approach of combining the interpretability of expert systems with the deductive power of data-driven ML can and should be transferred to other cases of clinical decision support.

Acknowledgments

Funding by the European Research Council (ERC), grant agreement No 647573, from German Research Foundation, GRK 2297 MathCoRe, and from the Klaus-Tschira-Foundation via Informatics for Life are gratefully acknowledged.

References

[1] A. Mincholé, B. Rodriguez, Artificial intelligence for the electrocardiogram, Nature Medicine 25 (1) (2019) 22.

[2] A. Vaish, P. Kumari, A comparative study on machine learning algorithms in emotion state recognition using ECG, in: Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012, Springer, 2014, pp. 1467–1476.

[3] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, A. Y. Ng, Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network, Nature Medicine 25 (1) (2019) 65.

[4] Z. I. Attia, S. Kapa, F. Lopez-Jimenez, P. M. McKie, D. J. Ladewig, G. Satam, P. A. Pellikka, M. Enriquez-Sarano, P. A. Noseworthy, T. M. Munger, et al., Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram, Nature Medicine 25 (1) (2019) 70.

[5] I. Fernández-Ruiz, Artificial intelligence to improve the diagnosis of cardiovascular diseases, Nature Reviews Cardiology (2019) 1.

[6] H. Li, D. Yuan, X. Ma, D. Cui, L. Cao, Genetic algorithm for the optimization of features and neural networks in ECG signals classification, Scientific reports 7 (2017) 41011.

[7] R. De Ponti, R. Marazzi, L. Zoli, F. Caravati, S. Ghiringhelli, J. A. Salerno-Uriarte, Electroanatomic mapping and ablation of macroreentrant atrial tachycardia: comparison between successfully and unsuccessfully treated cases, Journal of cardiovascular electrophysiology 21 (2) (2010) 155–162.

[8] J. M. Bumgarner, C. T. Lambert, A. A. Hussein, D. J. Cantillon, B. Baranowski, K. Wolski, B. D. Lindsay, O. M. Wazni, K. G. Tarakji, Smartwatch algorithm for automated detection of atrial fibrillation, Journal of the American College of Cardiology 71 (21) (2018) 2381–2388.

[9] Y. Guo, H. Wang, H. Zhang, T. Liu, Z. Liang, Y. Xia, L. Yan, Y. Xing, H. Shi, S. Li, et al., Mobile photoplethysmographic technology to detect atrial fibrillation, Journal of the American College of Cardiology 74 (19) (2019) 2365–2375.

[10] K. Elkholey, M. M. Lofgren, K. Q. Meeks, Z. U. A. Asad, B. Freedman, S. Stavrakis, Screening for atrial fibrillation in native americans using smartphone-based ecg, Circulation 140 (Suppl_1) (2019) A13895–A13895.

[11] M. Mutke, N. Brasier, C. Raichle, M. Doerr, J. Eckstein, C. C. research group, P1938 comparing atrial fibrillation detection algorithms in smart devices on validated mobile ecg data, European Heart Journal 39 (suppl_1) (2018) ehy565–P1938.

[12] A. Shiyovich, A. Wolak, L. Yacobovich, A. Grosbard, A. Katz, Accuracy of diagnosing atrial flutter and atrial fibrillation from a surface electrocardiogram by hospital physicians: Analysis of data from internal medicine departments, The American Journal of the Medical Sciences 340 (4) (2010) 271–275.

[13] B. Knight, G. Michaud, S. Strickberger, F. Morady, Electrocardiographic differentiation of atrial flutter from atrial fibrillation by physicians, Journal of Electrocardiology 32 (1999) 315–319.

[14] D. Krummen, M. Patel, H. Ngyen, G. Ho, D. Kazi, P. Clopton, M. Holland, S. Greenberg, G. Feld, M. Faddis, S. Narayan, Accurate ECG diagnosis of atrial tachyarrhythmias using quantitative analysis: A prospective diagnostic and cost-effectiveness study, Journal of Cardiovascular Electrophysiology 21 (2010) 1251–1259.

[15] P. Kirchhof, S. Benussi, D. Kotecha, A. Ahlsson, D. Atar, B. Casadei, M. Castella, H.-C. Diener, H. Heidbuchel, J. Hendriks, et al., 2016 ESC

Guidelines for the management of atrial fibrillation developed in collaboration with EACTS, European Heart Journal 37 (38) (2016) 2893–2962. arXiv:http://oup.prod.sis.lan/eurheartj/article-pdf/37/38/2893/23787249/ehw210.pdf, doi:10.1093/eurheartj/ehw210.

[16] N. Sawhney, R. Anousheh, W. Chen, G. K. Feld, Circumferential pulmonary vein ablation with additional linear ablation results in an increased incidence of left atrial flutter compared with segmental pulmonary vein isolation as an initial approach to ablation of paroxysmal atrial fibrillation, Circulation: Arrhythmia and Electrophysiology 3 (3) (2010) 243–248.

[17] E. Scholz, F. Kehrle, S. Vossel, A. Hess, E. Zitron, H. Katus, S. Sager, Discriminating atrial flutter from atrial fibrillation using a multilevel model of atrioventricular conduction, Heart Rhythm 11 (5) (2014) 877–884. URL https://mathopt.de/PUBLICATIONS/Scholz2014.pdf

[18] G. E. Box, D. R. Cox, An analysis of transformations, Journal of the Royal Statistical Society: Series B (Methodological) 26 (2) (1964) 211–243.

[19] I.-K. Yeo, R. A. Johnson, A new family of power transformations to improve normality or symmetry, Biometrika 87 (4) (2000) 954–959.

[20] J. Dougherty, R. Kohavi, M. Sahami, Supervised and unsupervised discretization of continuous features, in: Machine Learning 1995 Proceedings, Elsevier, 1995, pp. 194–202.

[21] S. Garcia, J. Luengo, J. A. Sáez, V. Lopez, F. Herrera, A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning, IEEE Transactions on Knowledge and Data Engineering 25 (4) (2012) 734–750.

[22] P. A. Estévez, M. Tesmer, C. A. Perez, J. M. Zurada, Normalized mutual information feature selection, IEEE Transactions on Neural Networks 20 (2) (2009) 189–201.

[23] G. Katz, E. C. R. Shin, D. Song, ExploreKit: Automatic feature generation and selection, in: 2016 IEEE 16th International Conference on Data Mining (ICDM), IEEE, 2016, pp. 979–984.

[24] T. Bikmukhametov, J. Jäschke, Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models, Computers & Chemical Engineering 138 (2020) 106834. doi:10.1016/j.compchemeng.2020.106834. URL http://www.sciencedirect.com/science/article/pii/S0098135419313675

[25] C. Rackauckas, Y. Ma, J. Martensen, C. Warner, K. Zubov, R. Supekar, D. Skinner, A. Ramadhan, Universal differential equations for scientific machine learning, arXiv preprint arXiv:2001.04385.

[26] A. Heinlein, A. Klawonn, M. Lanser, J. Weber, Machine learning in adaptive domain decomposition methods—predicting the geometric location of constraints, SIAM Journal on Scientific Computing 41 (6) (2019) A3887–A3912. doi:10.1137/18M1205364.

[27] M. Raissi, G. E. Karniadakis, Hidden physics models: Machine learning of nonlinear partial differential equations, Journal of Computational Physics 357 (2018) 125 – 141. doi:10.1016/j.jcp.2017.11.039.
URL http://www.sciencedirect.com/science/article/pii/S0021999117309014

[28] M. Raissi, P. Perdikaris, G. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, Journal of Computational Physics 378 (2019) 686 – 707. doi:10.1016/j.jcp.2018.10.045.
URL http://www.sciencedirect.com/science/article/pii/S0021999118307125

[29] S. Yan, Y. He, T. Tang, T. Wang, Drag coefficient prediction for non-spherical particles in dense gas–solid two-phase flow using artificial neural network, Powder Technology 354 (2019) 115–124.

[30] E. Qian, B. Kramer, B. Peherstorfer, K. Willcox, Lift & learn: Physics-informed machine learning for large-scale nonlinear dynamical systems, Physica D: Nonlinear Phenomena 406 (2020) 132401. doi:10.1016/j.physd.2020.132401.
URL http://www.sciencedirect.com/science/article/pii/S0167278919307651

[31] A. Yazdani, M. Raissi, G. E. Karniadakis, Systems biology informed deep learning for inferring parameters and hidden dynamics, bioRxiv (2019) 865063.

[32] G. Kissas, Y. Yang, E. Hwuang, W. R. Witschey, J. A. Detre, P. Perdikaris, Machine learning in cardiovascular flows modeling: Predicting arterial blood pressure from non-invasive 4d flow mri data using physics-informed neural networks, Computer Methods in Applied Mechanics and Engineering 358 (2020) 112623. doi:10.1016/j.cma.2019.112623.
URL http://www.sciencedirect.com/science/article/pii/S0045782519305055

[33] A. L. Hodgkin, A. F. Huxley, A quantitative description of membrane current and its application to conduction and excitation in nerve, The Journal of Physiology 117 (1952) 500–544.

[34] A. F. Villaverde, J. R. Banga, Structural properties of dynamic systems biology models: Identifiability, reachability, and initial conditions, Processes 5 (2). doi:10.3390/pr5020029.
URL https://www.mdpi.com/2227-9717/5/2/29

[35] T. Mazgalev, P. J. Tchou, Atrial-AV Nodal Electrophysiology: A View from the Millennium, Wiley, 2000.

[36] Y. Watanabe, L. Dreifus, Second degree atrioventricular block, Cardiovascular Research 1 (1967) 150–158.

[37] B. Kosowsky, P. Latif, A. Radoff, Multilevel atrioventricular block, Circulation 54 (1976) 914–921.

[38] R. Slama, J. F. Leclercq, M. Rosengarten, P. Coumel, Y. Bouvrain, Multilevel block in the atrioventricular node during atrial tachycardia and flutter alternating with wenckebach phenomenon, British Heart Journal 42 (4) (1979) 463–470.

[39] L. Littmann, R. H. Svenson, Atrioventricular alternating Wenckebach periodicity: Conduction patterns in multilevel block, The American Journal of Cardiology 49 (4) (1982) 855–862.

[40] A. Castellanos, J. Diaz, A. Interian, R. J. Myerburg, Wenckebach's periods or alternating wenckebach's periods during 4:1 atrioventricular block?, Journal of Electrocardiology 38 (2005) 157–159.

[41] O. Wolkenhauer, C. Auffray, O. Brass, J. Clairambault, A. Deutsch, D. Drasdo, F. Gervasio, L. Preziosi, P. Maini, A. Marciniak-Czochra, et al., Enabling multiscale modeling in systems medicine, Genome medicine 6 (3) (2014) 21.

[42] F. Fröhlich, F. J. Theis, J. O. Rädler, J. Hasenauer, Parameter estimation for dynamical systems with discrete events and logical operations, Bioinformatics 33 (7) (2017) 1049–1056.

[43] A. Kremling, J. Geiselmann, D. Ropers, H. de Jong, An ensemble of mathematical models showing diauxic growth behaviour, BMC systems biology 12 (1) (2018) 1–16.

[44] A. Tsipa, J. A. Pitt, J. R. Banga, A. Mantalaris, A dual-parameter identification approach for data-based predictive modeling of hybrid gene regulatory network-growth kinetics in pseudomonas putida mt-2, Bioprocess and biosystems engineering.

[45] M. E. Josephson, Clinical Cardiac Electrophysiology: Techniques and Interpretations, 4th Edition, Lippincott Williams & Wilkins, 2008.

[46] K. F. Wenckebach, H. Winterberg, Die unregelmäßige Herztätigkeit, Wilhelm Engelnmann, 1927.

[47] P. Denes, L. Levy, A. Pick, K. M. Rosen, The incidence of typical and atypical A-V Wenckebach periodicity, American Heart Journal 89 (1) (1975) 26–31.

[48] D. H. Spodick, Seven-cycle wenckebach period without atypical features, American Heart Hospital Journal 2 (1) (2004) 64.

[49] H. S. Friedman, J. A. C. Gomes, J. I. Haft, An analysis of Wenckebach periodicity, Journal of Electrocardiology 8 (4) (1975) 307–315.

[50] J. Hay, Bradycardia and cardiac arrhythmias produced by depression of certain functions of the heart, Lancet 1 (1906) 138–143.

[51] S. S. Barold, B. Lüderitz, John hay and the Earliest Description of Type II Second-Degree Atrioventricular Block, The American Journal of Cardiology 87 (12) (2001) 1433–1435.

[52] E. O. R. de Medina, R. Bernard, P. Coumel, A. Damato, C. Fisch, D. Krikler, N. Mazur, F. Meijler, L. Mogensen, P. Moret, Z. Pisa, H. Wellens, Who/isc task force. Definition of terms related to cardiac rhythm, American Heart Journal 95 (1978) 796–806.

[53] B. Surawicz, H. Uhley, R. Borun, M. Laks, L. Crevasse, K. Rosen, W. Nelson, W. Mandel, P. Lawrence, L. Jackson, N. Flowers, J. Clifton, J. Greenfield, E. R. D. Medina, The quest for optimal standardization of terminology and interpretation, American Heart Journal 41 (1) (1978) 130–145.

[54] D. P. Zipes, J. P. Dimarco, P. C. Gillette, W. M. Jackman, R. J. Myerburg, S. H. Rahimtoola, J. L. Ritchie, M. D. Cheitlin, A. Garson, R. J. Gibbons, R. P. Lewis, R. A. O'Rourke, T. J. Ryan, R. C. Schlant, Guidelines for clinical intracardiac electrophysiological and catheter ablation procedures, Journal of the American College of Cardiology 26 (2) (1995) 555–573.

[55] S. S. Barold, 2:1 Atrioventricular block: Order from chaos, The American Journal of Emergency Medicine 19 (3) (2001) 214–217.

[56] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (2011) 27:1–27:27.

[57] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (3) (1995) 273–297. `doi:10.1007/BF00994018`.

[58] A. Camm, P. Kirchhof, G. L. et al., Guidelines for the management of atrial fibrillation, European Heart Journal 31 (2010) 2369–2429.

[59] F. Kehrle, Inverse simulation for cardiac arrhythmia, Ph.D. thesis, Otto-von-Guericke University Magdeburg (2018).
URL `https://mathopt.de/PUBLICATIONS/Kehrle2018.pdf`

[60] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. C. Ivanov, R. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, H. E. Stanley, Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals, Circulation [Online] 101 (23) (2000) e215–e220.
URL `https://physionet.org/content/ahadb/1.0.0`

[61] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, CoRR abs/1412.6980. arXiv:1412.6980.
URL http://arxiv.org/abs/1412.6980

[62] B. L. Hoppe, A. Kahn, G. Feld, A. Hassankhani, S. Narayan, Separating atrial flutter from atrial fibrillation with apparent electrocardiographic organization using dominant and narrow F-wave spectra, Journal of the American College of Cardiology 46 (2005) 2079–2087.

[63] F. Bogun, D. Anh, G. Kalahasty, E. Wissner, C. B. Serhal, R. Bazzi, W. Weaver, C. Schuger, Misdiagnosis of atrial fibrillation and its clinical consequences, The American Journal of the Medical Sciences 117 (9) (2004) 636–642.

[64] K. Kettering, V. Dörnberger, R. Lang, R. Vonthein, R. Suchalla, R. F. Bosch, C. Mewis, B. Eigenberger, V. Kühlkamp, Enhanced detection criteria in implantable cardioverter defibrillators: Sensitivity and specificity of the stability algorithm at different heart rates, Pacing and Clinical Electrophysiology 24 (2001) 1325–1333.

[65] E. Strickland, Ibm watson, heal thyself: How ibm overpromised and underdelivered on ai health care, IEEE Spectrum 56 (04) (2019) 24–31.

[66] C. Combi, G. Pozzi, Clinical information systems and artificial intelligence: Recent research trends, Yearbook of medical informatics 28 (1) (2019) 83.

[67] N. Baker, F. Alexander, T. Bremer, A. Hagberg, Y. Kevrekidis, H. Najm, M. Parashar, A. Patra, J. Sethian, S. Wild, et al., Workshop report on basic research needs for scientific machine learning: Core technologies for artificial intelligence, Tech. rep., USDOE Office of Science (SC), Washington, DC (United States) (2019).

[68] T. Syeda-Mahmood, Role of big data and machine learning in diagnostic decision support in radiology, Journal of the American College of Radiology 15 (3) (2018) 569–576.

[69] E. J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nature Medicine 25 (1) (2019) 44–56. doi:10.1038/s41591-018-0300-7.

[70] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, K. Zhang, The practical implementation of artificial intelligence technologies in medicine, Nature Medicine 25 (1) (2019) 30–36. doi:10.1038/s41591-018-0307-0.

[71] National Science and Technology Council, National artificial intelligence research and development strategic plan (2016).
URL https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/national_ai_rd_strategic_plan.pdf

[72] Bundesregierung der Bundesrepublik Deutschland, Nationale Strategie für künstliche Intelligenz (2018).
URL https://www.ki-strategie-deutschland.de/home.html

[73] S. Ahmed, A. Claughton, P. A. Gould, Atrial flutter - diagnosis, management and treatment, in: Abnormal Heart Rhythms, IntechOpen, 2015.

[74] F. Garcia-Cosio, A. P. Fuentes, A. N. Angulo, Clinical approach to atrial tachycardia and atrial flutter from an understanding of the mechanisms. electrophysiology based on anatomy, Revista Española de Cardiologia (English Edition) 65 (4) (2012) 363–375.

[75] K. Hornik, Approximation capabilities of multilayer feedforward networks, Neural networks 4 (2) (1991) 251–257.

[76] P. Kidger, T. Lyons, Universal approximation with deep narrow networks, in: Conference on Learning Theory, 2020, pp. 2306–2327.

[77] A. D. Jagtap, K. Kawaguchi, G. E. Karniadakis, Adaptive activation functions accelerate convergence in deep and physics-informed neural networks, Journal of Computational Physics 404 (2020) 109136. doi:10.1016/j.jcp.2019.109136.
URL http://www.sciencedirect.com/science/article/pii/S0021999119308411

[78] F. Jost, J. Zierk, T. T. Le, T. Raupach, J. Zierk, M. Rauh, M. Suttorp, M. Stanulla, M. Metzler, S. Sager, Model-based simulation of maintenance therapy of childhood acute lymphoblastic leukemia, Frontiers in Physiology 11 (2020) 217. doi:10.3389/fphys.2020.00217.
URL https://www.frontiersin.org/article/10.3389/fphys.2020.00217

[79] F. Jost, E. Schalk, D. Weber, H. Döhner, T. Fischer, S. Sager, Model-based optimal aml consolidation treatment, IEEE Transactions on Biomedical EngineeringAccepted.
URL https://arxiv.org/abs/1911.08980

[80] P. Lilienthal, M. Tetschke, E. Schalk, T. Fischer, S. Sager, Optimized and personalized phlebotomy schedules for patients suffering from polycythemia vera, Frontiers in PhysiologyAccepted.

[81] S. Sager, Optimization and clinical decision support, Optima 104 (2018) 1–8.
URL http://www.mathopt.org/Optima-Issues/optima104.pdf

[82] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215.