PLOS ONE

RESEARCH ARTICLE

Expert-enhanced machine learning for cardiac arrhythmia classification

Sebastian Sager[1,2]*, **Felix Bernhardt**[1], **Florian Kehrle**[2,3], **Maximilian Merkert**[4], **Andreas Potschka**[5], **Benjamin Meder**[2,3], **Hugo Katus**[2,3,6], **Eberhard Scholz**[2,7]

1 Department of Mathematics, Otto-von-Guericke University, Magdeburg, Germany, **2** Informatics for Life, Heidelberg, Germany, **3** Department of Internal Medicine III, University Hospital Heidelberg, Heidelberg, Germany, **4** Institute of Optimization, Technical University Braunschweig, Braunschweig, Germany, **5** Institute of Mathematics, Clausthal University of Technology, Clausthal-Zellerfeld, Germany, **6** German Centre for Cardiovascular Research, Heidelberg, Germany, **7** GRN Gesundheitszentren Rhein-Neckar gGmbH, Schwetzingen, Germany

* sager@ovgu.de

Citation: Sager S, Bernhardt F, Kehrle F, Merkert M, Potschka A, Meder B, et al. (2021) Expert-enhanced machine learning for cardiac arrhythmia classification. PLoS ONE 16(12): e0261571. https://doi.org/10.1371/journal.pone.0261571

Editor: Friedhelm Schwenker, Ulm University, GERMANY

Received: September 20, 2021

Accepted: December 5, 2021

Published: December 23, 2021

Data Availability Statement: All data and implementations are available from the Supporting information files.

Funding: Funding by the European Research Council (ERC), grant agreement No 647573, from German Research Foundation (Deutsche Forschungsgemeinschaft), GRK 2297 MathCoRe, and from the Klaus-Tschira-Foundation via Informatics for Life are gratefully acknowledged. Funders did not play any role in the study design, data collection and analysis, ecision to publish, or preparation of the manuscript.

Abstract

We propose a new method for the classification task of distinguishing atrial fibrillation (AFib) from regular atrial tachycardias including atrial flutter (AFlu) based on a surface electrocardiogram (ECG). Recently, many approaches for an automatic classification of cardiac arrhythmia were proposed and to our knowledge none of them can distinguish between these two. We discuss reasons why deep learning may 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 by solving an additional regression problem with complicated combinatorial substructures. The resultant can be seen as a novel machine learning model that incorporates expert knowledge on the pathophysiology of atrial flutter. Our approach achieves an unprecedented accuracy of 82.84% and an area under the receiver operating characteristic (ROC) curve of 0.9, which classifies as "excellent" according to the classification indicator of diagnostic tests. One additional advantage of our approach is the inherent interpretability of the classification results. Our features give insight into a possibly occurring multilevel atrioventricular blocking mechanism, which may improve treatment decisions beyond the classification itself. Our research ideally complements existing textbook cardiac arrhythmia classification methods, which cannot provide a classification for the important case of AFib↔AFlu. The main contribution is the successful use of a novel mathematical model for multilevel atrioventricular block and optimization-driven inverse simulation to enhance machine learning for classification of the arguably most difficult cases in cardiac arrhythmia. A tailored Branch-and-Bound algorithm was implemented for the domain knowledge part, while standard algorithms such as Adam could be used for training.

Competing interests: The authors declare that no competing interests exist.

Introduction

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 [2–5] were due to heart rate variability, the particular shape of the electrocardiogram (ECG) curve (including low voltage flutter waves that correspond to atrial polarizations), or a mix of both. Wavelets have been used to extract features automatically [6], but this approach is limited to easy classification cases and does not directly provide physiologically interpretable features. Usually, parameters such as atrial cycle length are not provided, although they may be relevant for treatment decisions [7].

Moreover, none of the surveyed studies addressed the especially difficult case of atrial fibrillation (AFib) versus regular atrial arrhythmias including atrial flutter and focal atrial tachycardias with irregular ventricular response (AFlu), summarized as AFib↔AFlu hereafter. It is either 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 grouped together in deep learning (*"The atrial fibrillation class combined atrial fibrillation and atrial flutter"* [3]) and algorithms based on heart rate variability for smartwatches [8]. Studies that explicitly address "detection of AFib" in the title [9–11] can only detect the grouped 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 rates of misdiagnosis and possible causes have been reported [12–14]. This is concerning, as different treatments (often antiarrhythmics in AFib versus a highly successful ablation therapy in AFlu) are implied by the diagnosis [15]. Diagnosing atypical forms of AFlu is becoming increasingly important in clinical practice due to complications of left atrial ablation procedures [16]. See Scholz et al. (*"Discriminating atrial flutter from atrial fibrillation using a multilevel model of atrioventricular conduction"*) [17] for a more detailed discussion. The poor quality of expert opinion due to the difficult discrimination poses a challenge to automated classification by supervised ML, which often uses it for labeling training samples [3–5]. We used an expert analysis based on intracardiac measurements, which is only available with invasive procedures, as our gold standard.

Interestingly, the case AFib↔AFlu seems to be difficult for deep learning approaches. As stated before, the differentiation between AFib and AFlu has been avoided in Hannun et al. (*"Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network"*) [3], where a deep convolutional net with 34 layers was trained using 91232 singe-lead ECGs. Moreover, our results show poor performance of neural-network-based approaches. We conjecture that this is due to the non-continuous nature of the underlying process, which contrasts to the approximation properties of deep neural networks and the relatively small size of the training set.

Complementing previous work in automatic arrhythmia classification

Fig 1 visualizes our workflow. Deep learning (DL) can robustly distinguish samples of either AFib or AFlu from sinus rhythm and twelve cardiac arrhythmias [3] with high accuracy. Other

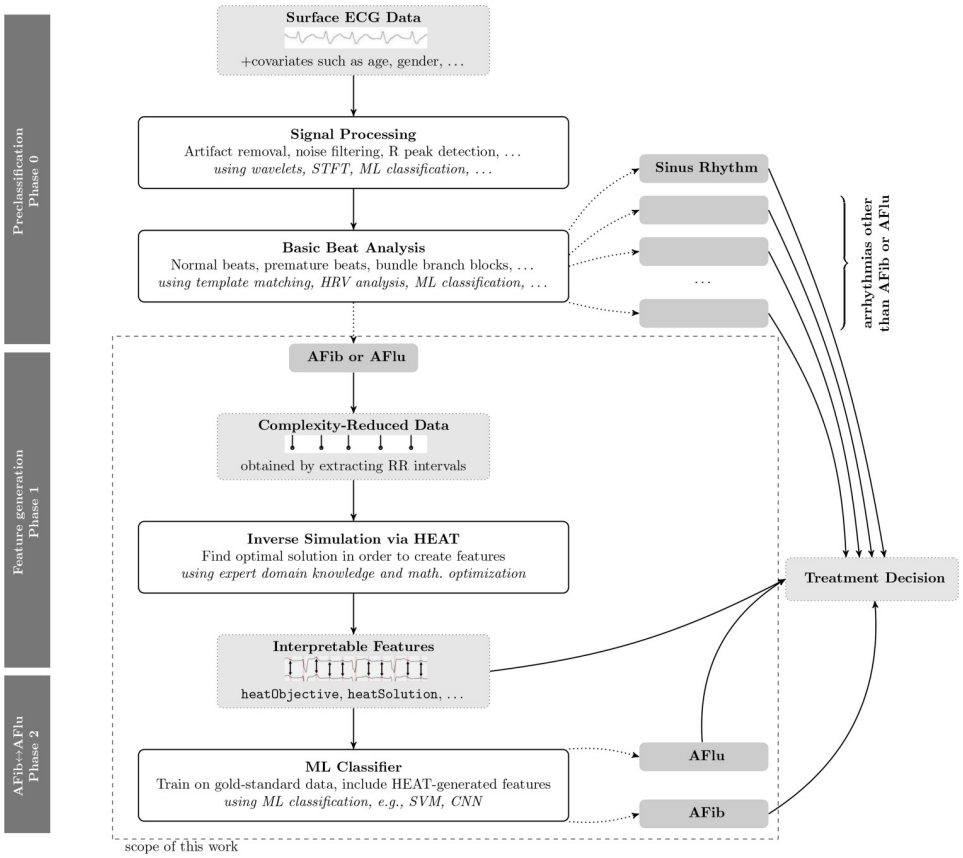

Fig 1. Visualization of our workflow from surface ECG to decision support for treatment. We focus on phases 1 (generation of physiologically interpretable features) and 2 (using them for AFib↔AFlu classification), thus assuming a pre-classification of all samples.

https://doi.org/10.1371/journal.pone.0261571.g001

studies achieved similar results [6, 9–11]. For a survey on general ECG-based automatic arrhythmia classification, see Luz et al. (*"ECG-based heartbeat classification for arrhythmia detection: A survey"*) [18].

As a reliable pre-classification (Phase 0) can thus be achieved, we focus here on Phase 1 (generation of physiologically interpretable features) and Phase 2 (using them for AFib↔AFlu classification). In the following, we assume that it has been verified that only either AFib or AFlu is present, which is also true for our gold standard data set (expert classification of intra-cardiac measurements that are only available after invasive procedures).

We propose to extend and complement the mentioned approaches with generated features based on a pathophysiological rationale allowing classification of AFib↔AFlu. Thus, our approach is not an alternative to previous work of automatic classification, but is rather complementary to it. In previous works, 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 any classification algorithm can be applied in Phase 2.

Feature generation and hybrid modeling

Feature construction has a long history, with early work dating back to the 1960s [19]. Since then, there has been a plethora of feature generation methods, such as polynomial [20], discretization [21, 22], normalization [23], or grouping operations involving min, max, averaging, etc. The current state-of-the-art in feature construction methods suffer 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 themselves can be based on machine learning technology [24], the difficulty to embed domain knowledge and to interpret the automatically generated and selected features still remains. Our proposed feature generation overcomes 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 experienced medical practitioners. in addition, exponential explosion of the feature set is not an issue because only a few additional real-valued features need to be added.

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

Combining machine learning models with domain knowledge is an active and promising field of research, e.g., [25, 26]. A survey on how first principle models can be combined in different ways with generic machine learning models is given in Bikmukhametov et al. (*"Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models"*) [27] in the context of process engineering systems. One way is to replace uncertain parts in differential equations with neural nets using the concept of universal differential equations [28]. ML can also be applied to make the solution of differential equations more efficient [29]. The alternative is to develop and use physics-informed or biology-informed machine learning approaches [30–34]. The general idea is to design ML 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 noisy clinical data in Kissas et al. (*"Machine learning in cardiovascular flows modeling: Predicting arterial blood pressure from non-invasive 4D flow MRI data using physics-informed neural networks"*) [35]. Here, arterial pressure was predicted from MRI data of blood velocity and wall displacement. Common results of these studies show that by combining physics-based and machine learning models it is possible to improve the performance of the purely black-box ML models making them more transparent and interpretable.

The mathematical model develop and applied in this study can be seen as a simplification of first-principle models for electrical conductivity in the heart, such as the Hodgkin–Huxley equations [36]. In this sense, our approach can also be interpreted as a biology-informed machine learning approach. See Villaverde et al. (*"Structural Properties of Dynamic Systems Biology Models: Identifiability, Reachability, and Initial Conditions"*) [37] for a survey of systems biology models and important properties.

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. As Douglas P. Zipes stated in 2000, the AV node is still *"a riddle wrapped in a mystery inside an enigma"* [38]. Key to solving this riddle is the idea of a multilevel AV block (MAVB) [39–43]. The tedious procedure of manually adjusting possible MAVB combinations has been successfully automated 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, 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 each training sample. The combination of a mathematical model and optimization algorithm could be seen as an interpretable expert system. The basic idea of using a mathematical model and inverse simulation for AFib↔AFlu classification has been published before in [17]. We report a significantly matured approach with a larger (4×) data set which allows for a systematic cross-validation, an improved mathematical model of MAVB with a better pathophysiological interpretation, a computational speed up to 5000×, and an increased accuracy (the area under the ROC of 0.9 in [17] was not cross-validated). 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 may overcome intrinsic approximation limits of deep learning for difficult-to-label and non-smooth systems that often occur in medicine and biology [44–47].

Structure of this paper

The paper is structured per PLOS One guidelines. In Section Methods 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 Results we present numerical results showing that the proposed approach reaches an unprecedented accuracy, while a direct use of neural networks perform poorly on the data. In Section Discussion 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 Conclusions.

Methods

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, the *absolute refractory period* (ARP) describes the duration in which a cell cannot be stimulated under any circumstances. The *relative refractory period* (RRP) describes the duration in which the cells can be stimulated under certain conditions, but may react with a modified conduction [48]. Depending on incoming signal and RRP, a block ratio of $n + 1$: n can occur, where $n+1$ is the number of incoming signals, and n the number of conducted signals. This ratio may vary due to changes in cell fatigue or in the frequency of the incoming signals, even on short time horizons. For larger values of n the conduction times may change as well.

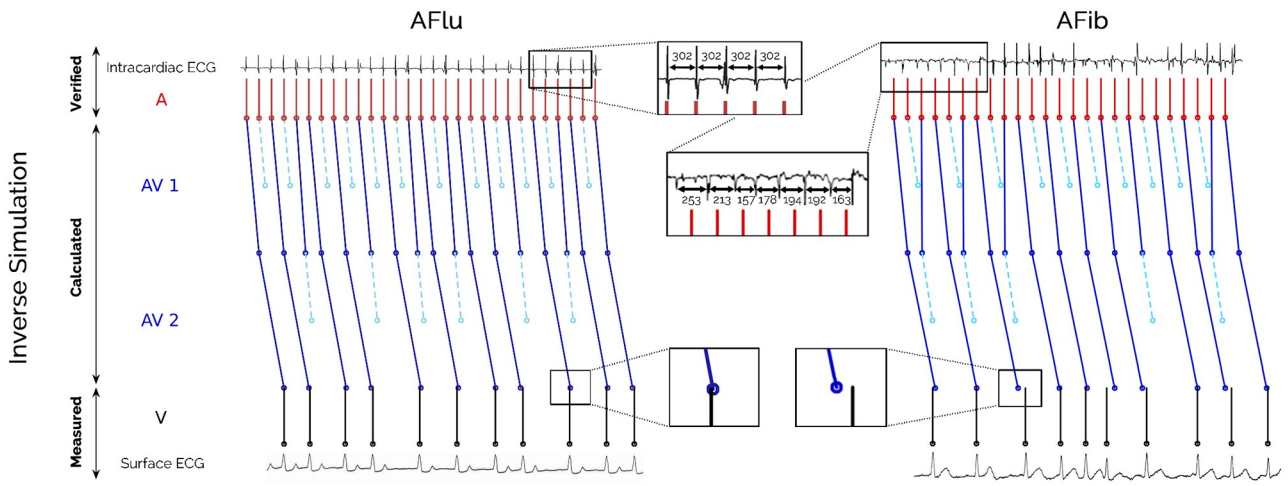

Fig 2. Visualization of our inverse simulation approach applied to samples of atrial flutter (AFlu, left, regular intracardiac measurement) versus atrial fibrillation (AFib, right, irregular intracardiac measurement) based on the surface electrocardiogram (ECG, bottom). In this example, a two-level atrioventricular (AV) block was calculated for both samples.

https://doi.org/10.1371/journal.pone.0261571.g002

Reviewing 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 (see Fig 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, allowing the measurements to be used 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. 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 chaotic behavior which cannot be explained well by the model (AFib). Comparing bottom zooms in Fig 2, cf. Scholz2014, it visually confirmes that for AFlu the calculated Δa corresponds well to the intracardiac measurements.

This theoretical concept allows to combine different blocking ratios $n + 1$: n on an unlimited number of levels. Possibly varying and linearly changing conduction times due to RRP are denoted as *Type I*. Sensibly, the number of possible combinations should be limited to avoid overfitting, reduce computational time, and stay close to clinical observations. We restricted our MAVB model to the five combinations shown in Fig 3 with a maximum of three block levels, consistent with cases described in recent publications.

The resulting mathematical model is a combination of most different classical and advanced block types, particularly, typical Type I block [49–51], atypical Type I block [50, 52], the special cases of 2:1 and 3:2 Type I blocks, Type II block [53–56], advanced second-degree AV Block [57, 58], and MAVB [39–43]. Invoking Occam's razor, this unified model also allows an efficient calculation of the most likely block for given RR data.

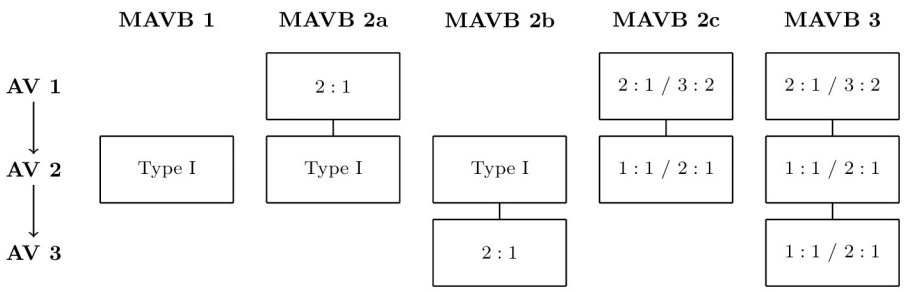

Fig 3. The five considered blocktypes, having up to three multilevel atrioventricular block (MAVB) levels.

https://doi.org/10.1371/journal.pone.0261571.g003

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. Internally, time points t_{ij} are calculated and denote, when the signal originating from signal j in the atrium reaches level i. Due to the assumed regularity in the atrium we have

$$t_{0j} \coloneqq t_s + j\Delta a$$

with an unknown offset t_s. On levels 1, 2, and 3 the equations for t_{ij} depend on the particular blocking type bt, and hence more complicated case differentiations: if the signal can be conducted,

$$t_{ij} \coloneqq t_{i-1,j} + f(oc)$$

with a linear function f depending on parameters oc, otherwise it will be blocked and can not be considered in the objective function. Details can be found in the PhD thesis [59] and in the survey paper [60]. The objective function is denoted by F_i where $F_i(x)$ measures the deviation of the resulting 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^*, particular values for Δa_i^*, bt_i^*, and oc_i^* that 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 (Δa,bt, oc) with lower and upper bounds for (Δa, oc) and five most clinically observed blocktypes of MAVB (see Fig 3). The bounds on the atrial cycle length Δa were determined using physiological observations [48] (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 by request.

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`).

Increasing accuracy and stability, we applied a moving horizon strategy to generate additional features. From the $n_{RR} = 22$ time intervals, we considered only $n_{sub} \in \mathcal{I} := \{10, \ldots, n_{RR}\}$ on windows $[1, 2, \ldots, n_{sub}]$ until $[n_{RR} - n_{sub} + 1, 2, \ldots, n_{RR}]$. This results in additional solutions $F_{i,n_{sub}}(x^*_{i,n_{sub}})$ for $i \in \mathcal{I}$. Investigating the robustness of solutions, we evaluated $F_{i,j}(x^*_{i,k})$ for $j, k \in \mathcal{I}$, the performance of the optimal solutions on time window j on time window k. We computed the features `HEATobj` and `HEATsol` for each subwindow of RR intervals. The moving horizon approach enabled us to compare of the HEAT simulation based on one time window with the raw RR intervals of a different one, as described above. We refer to the resulting time series of $n_{RR} - n_{sub} + 1$ 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.

Machine learning models

We used two classes of standard ML classification models: SVM and convolutional neural networks (CNN).

As SVM does not incorporate the temporal connection between sequential data, we first computed general features based on subsequences (N-Gram s) 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 $\leq n_{RR}$ 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. Prior to training use, each feature was standardized to zero mean and unit standard deviation. The necessary parameters for this transformation were computed on the training set and used for the model evaluation. Based on these features, we implemented a SVM model in scikit-learn based on the LIBSVM library [61]. The underlying model is described in Cortes et al. ("*Support-vector networks*") [62]. 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 $n_{sub} \in \{10, \ldots, 22\}$ were tuned as hyperparameters using grid search cross-validation.

We used a CNN architecture consisting of two convolutional blocks followed by one fully connected layer with rectified linear unit (ReLU) activation functions and one final fully connected layer with a sigmoid activation function and output dimension one. Each of the convolutional blocks consisted of two convolutional layers with ReLU activation functions and five filters of width two 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 eliminated as they gave no additional insight.

Table 1 shows the number of optimization parameters, scaling factors, and hyperparameters for the different approaches. The number of optimized parameters depends on the hyperparameter n_{sub} (the length of analyzed subsequences); therefore, 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.

Data

Our data set `heatDS` is a superset of one used in a previous study [17], which contains details of the data obtained from patients exhibiting AFib or AFlu with irregular ventricular response

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

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

https://doi.org/10.1371/journal.pone.0261571.t001

during invasive electrophysiological testing or catheter ablation. The retrospective data were extended to the period between 2011 and 2018 and 159 patients.

Classification AFib↔AFlu was performed using electrical signals measured at the atrial electrodes by an expert in the field of cardiac electrophysiology for all 159 patients. 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, referenced in the European guideline for the management of AFib [63]. 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.

We hypothesized that the dynamics of ventricular activations in short time periods contain enough information for 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, with a precision of 1 ms. Segments containing premature ventricular beats were excluded, which can be easily recognized by physicians or algorithms in clinical practice.

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

In Kehrle (*"Inverse Simulation for Cardiac Arrhythmia"*) [59], we validated a previous version of our algorithm against other, smaller data sets from the publications focused on AFib↔AFlu discrimination. Unfortunately, there are no larger data sets available that can be used as an extended benchmark. Usually, these do not differentiate between AFib and AFlu specifically, or they do not classify supraventricular tachycardias at all, such as the American Heart Association ECG Database for example [64]. Therefore, all of the data in studies [8–11] could not be used, as it is unlabeled with respect to AFib↔AFlu.

Table 2. Average accuracies and areas under receiver operating characteristic (ROC) curve with standard deviations for the different approaches.

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

https://doi.org/10.1371/journal.pone.0261571.t002

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 2.40 GHz 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 [65] with default parameters. The computational times were roughly 20 ms per HEAT call (times 380 samples times number of considered subproblems per sample), 30 min for training SVM, and 3 d for training CNN.

Results

Accuracies for different feature sets and ML models

We show the mean accuracies 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 Features and feature sets and Machine learning models.

When directly applied to the input data of upto 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 Fig 6.

For an implementation of a 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 than that 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 considerably stable and no significant differences occurred for different kernel types. The approach to pre-process `rawRR` using medical expert knowledge (HEAT) can be seen as an approach that increases sensitivity without overfitting the ML model.

Interpretability

We observed that the calculated objective function values $F_i(x_i^*)$ were the most decisive feature for classification, and the features associated with x_i^* are interesting for clinical interpretation. Fig 4 shows how knowing the atrial cycle length Δa^* may be helpful for an a-posteriori

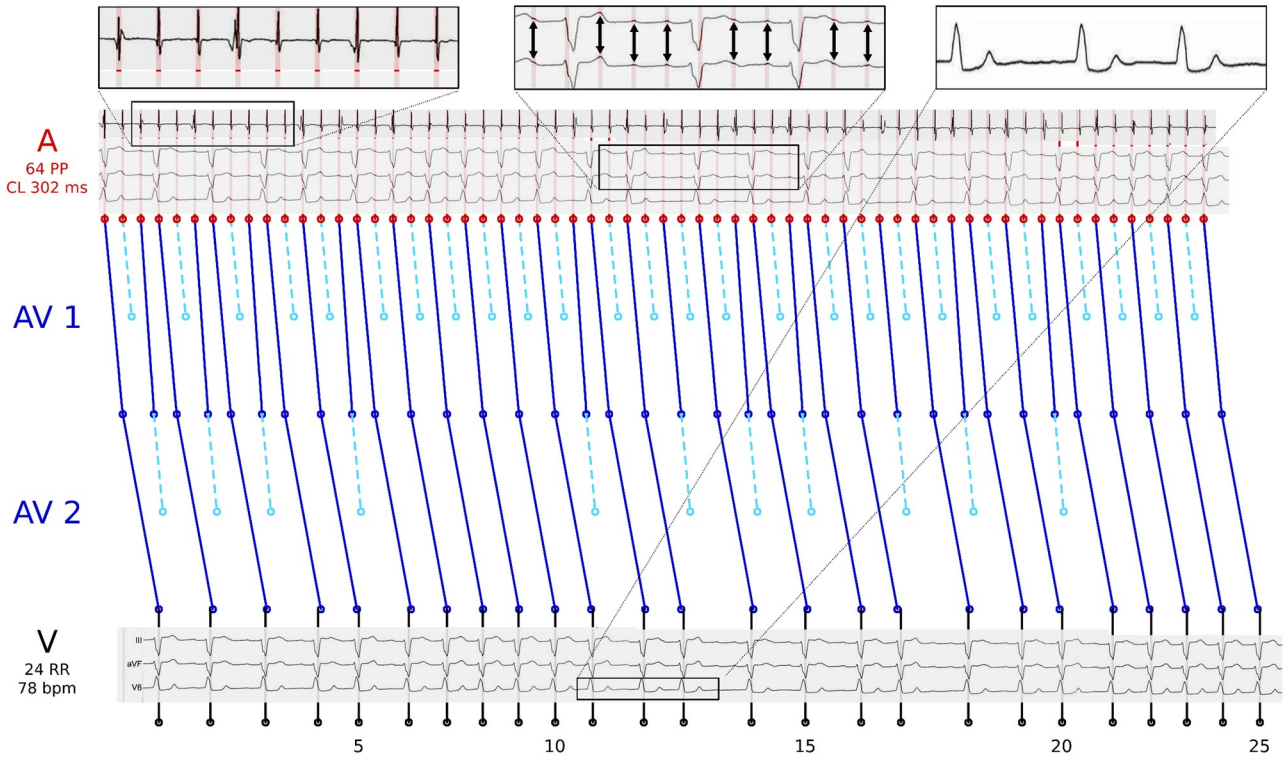

Fig 4. Exemplary illustration of how the feature *atrial cycle length* derived from a HEAT solution can be a posteriori pathophysiologically interpreted and used.

https://doi.org/10.1371/journal.pone.0261571.g004

identification of flutter waves for AFlu in a surface ECG. The figure shows observed and simulated data, as in left-hand side of Fig 2, 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 can help the physician, when reanalyzing the ECG. An overlay of Δa makes spotting atrial activations in the surface ECG easier (middle zoom).

Fig 5 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 Fig 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 Figs 4 and 5 with two and three levels with varying blockings, respectively, give insight into the pathophysiology of the AV node and may be useful for treatment planning.

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.

Moving horizon approach

The results in Table 2 seem to indicate that additional accuracy can be obtained using the feature `HEATseries`. It consists of time-series data generated from several calls to HEAT for

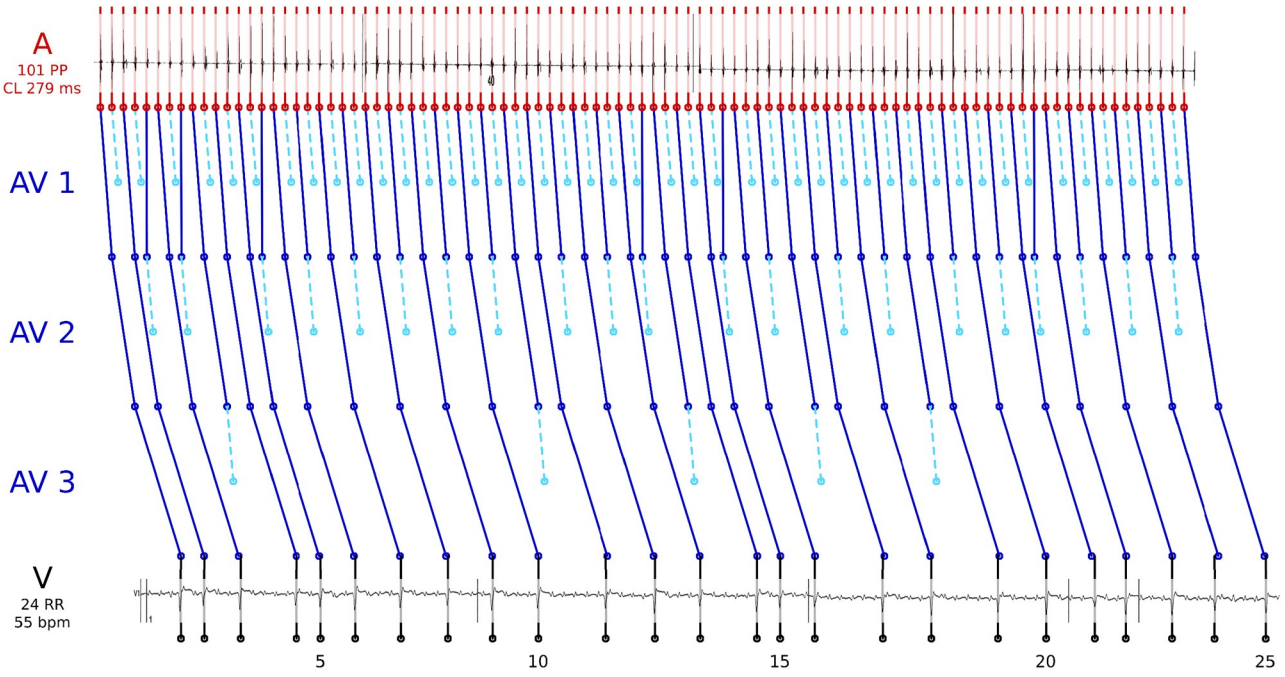

Fig 5. Exemplary illustration of how the feature *blocktype* derived from a HEAT solution can be a posteriori pathophysiologically interpreted and used.

https://doi.org/10.1371/journal.pone.0261571.g005

input data obtained from a moving horizon approach. As explained above, $n_{sub} \in \{10, \ldots, n_{RR}\}$ was optimized as a hyperparameter, with $n_{sub} = 17$ giving the best results. The overall number of time intervals $n_{RR} = 22$ was fixed. Therefore, the time series in `HEATseries` corresponded to entries for six different optimization problems (1 . . .17 to 6 . . .22).

An interesting and promising question is regarding 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 was presented after data from many patients with small numbers of RR intervals were already collected. 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. A study showed large potential with accuracy increasing from 82.94% to 92.50% for long time horizons of $n_{RR} = 90$ intervals. However, this result needs to be cross-validated on larger data sets.

Discussion

Impact, accuracy, and applicability

Being able to classify AFib↔AFlu is clinically relevant. There are a variety of treatments (anti-arrhythmics, various ablations and ablation systems) with different side effects and cure rates. A correct classification is imperative to choose the best treatment [15]. Therefore, use of the proposed approach for clinical decision support may be helpful, especially when considering the excellent classification accuracy and interpretability of calculated features and the difficulty of the classification task for unexperienced clinicians.

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 [12–14] and was explicitly excluded in recent studies [3]. AFib may be overdiagnosed because of coarse fibrillatory waves, which are reminiscent of AFlu [13, 66], the presence of artifacts, or premature atrial complexes [67]. AFlu may be overdiagnosed because the low-voltage flutter waves that indicate AFib are barely discernible in the surface ECG (compare Figs 2 and 4), or because a pseudo-regularization may occur [68] (see Section Classification failures). 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 is currently untreated by them.

Using HEAT for an a-priori calculation of `heatObjective` was significantly more successful with an average accuracy of 77.58%, even though 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.

Using ML with HEAT-generated features has the drawback; Each classification sample requires calculating the optimal solution of the MAVB. However, the additional 20 ms should be acceptable in a clinical context and negated by several advantages:

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

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

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

Interpretability

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

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 assessible to humans. HEAT provides `HEATsol`, the optimal solution $x^* = (\Delta a^*, bt^*, oc^*)$. These values can be interpreted by experts, and used for maiking treatment decisions. For example, the atrial cycle length Δa^* proposed by HEAT can help the physicians when reanalyzing the ECG (compare Fig 4). Furthermore, the absolute

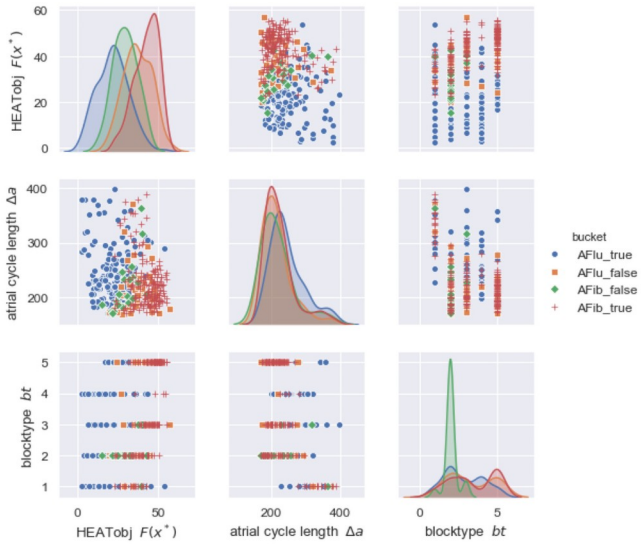

Fig 6. Representative pairwise plot of features obtained from a `heatSolution` SVM classification, compare Table 2.

https://doi.org/10.1371/journal.pone.0261571.g006

cycle length can 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* [69]. 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* [70]. We believe that estimates of the atrial cycle length or the blocktype (compare Figs 4 and 5) can be a valuable asset to clinical decision making.

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 considered, SVM outperformed our CNN architectures as discussed in the next subsection. We expect a different behavior if neural network architectures are 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 Multilevel atrioventricular block (MAVB). This generic approach is preferable for the aforementioned reason of interpretability and it obsoletes 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 a few model parameters that are calculated along with the objective function value. The latter alone was decisive and was

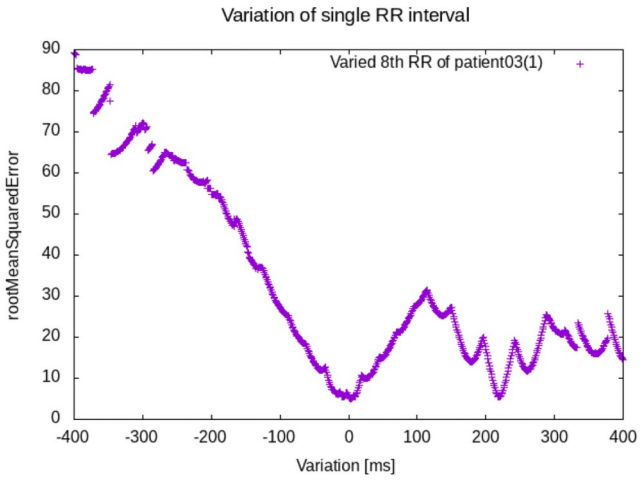

Fig 7. Fluctuation of the objective function of our mathematical model with respect to shifts in one input signal of a data sample.

https://doi.org/10.1371/journal.pone.0261571.g007

enough for a high-accuracy one-dimensional linear classifier, using a simple threshold value compare the entry for `heatObjective` in Table 2. The additional features considered in `heatSolution` increased accuracy, although we see the main benefit of block type, atrial cycle length, and conduction constants in the physiological interpretability. Future work should focus on consideration of sets of optimal solutions and solutions on moving time horizons. In this context, the impact of `heatSolution` may improve.

Approximation properties of ML 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 [71] or the number of layers for a fixed number of neurons per layer [72] increase. However, it is also well known that these theoretical results are obtained at the price of a potentially large number of weights distributed over the hidden layers of the neural net. Adaptive activation functions have better approximation properties [73], but the main difficulty of current architectures is the same. To get an idea why CNNs do not perform well on AFib↔AFlu, for deep nets with 34 layers as in [3] as well as in our prototypical implementation, we analyze Fig 7.

Fig 7 shows the feature `HEATobj`, 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. Sixteen of them are kept fixed, while one particular interval length in the middle was varied with deviations of -400 ms to +400 ms in intervals of 1 ms. The plot shows locally quadratic behavior, due to the quadratic objective function (Euclidean norm). The discontinuities are due to the clipping of solutions that result in deviations of more than 150 ms between signals. The main takeaway 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 difficulty for this

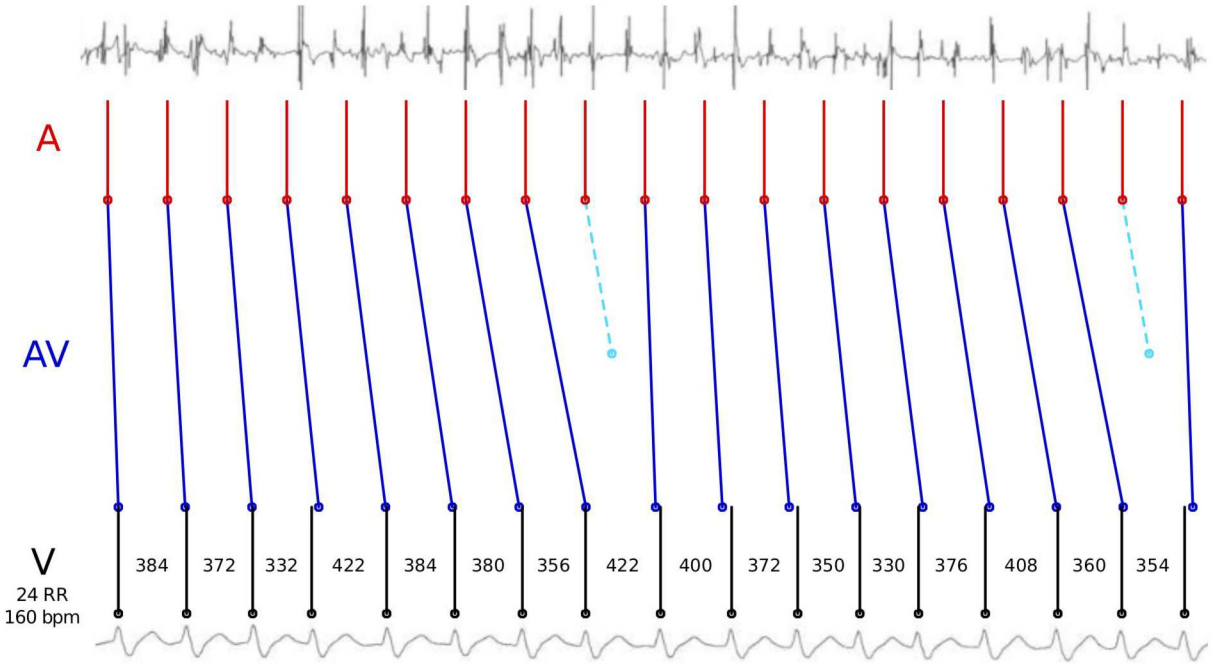

Fig 8. Example of an atrial fibrillation input that is misclassified due to pseudo-regularization.

https://doi.org/10.1371/journal.pone.0261571.g008

classification task, only a few labeled training data sets are available. We conjecture that it will be difficult to train CNNs with a reasonable classification accuracy without using domain knowledge.

Classification failures

While our novel approach resulted in excellent area under the curve values, there were still misclassification samples. Fig 8 shows an atrial fibrillation case with a very fast (160 beats per minute), but pseudo-regular ventricular contraction, shown in the surface lead at the bottom. The atrial contraction, however, is totally chaotic as shown by intracardiac measurements displayed in the top. Due to this pseudo-regularization, the best MAVB simulation matched the observed data considerably well and led to a misclassification. It is well known that at very high frequencies of AFib, a pseudo-regularization can occur [68]. 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 may be more likely to match a regular MAVB or even a 1: 1 conduction. In our approach, pseudo-regularizations result in relatively low objective function values which impair correct classification.

Just as for experts, the presence of artifacts or premature atrial complexes [67] may lead to a misclassification. It is an open question how to extend the mathematical model in Section Multilevel atrioventricular block (MAVB) for automatic detection of pseudo-regularization and increased specifity without impairing sensitivity. Using the feature *atrial cycle length* more elaborately or additionally classifying the flutter waves may be helpful in this context.

An intrinsic limitation for classification accuracy using our approach arises from false positives, cases of AFib that "by chance" are very close to multilevel blocks. The mathematical

question of how dense random `rawRR` instances are in the space of all MAVB solutions is open.

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 be interpreted by clinicians. Diagnosis of other cardiac arrhythmias could be done in a similar way. For diseases such as acute leukemias [74, 75] or polycythemia vera [76], there are mathematical models that have been validated with measurement data, and contain estimated personalized model parameters like stem cell proliferation rates. Such hidden parameters usually cannot be observed directly and can be very useful for clinical decision-making [60].

Our interdisciplinary approach with cardiologists and mathematical optimizers has several obvious benefits [77]. One of them is that the role of HEAT can be seen as a well-informed agent interacting with a surrounding machine learning environment. Such an approach was introduced and discussed in Holzinger (*"Interactive machine learning for health informatics: when do we need the human-in-the-loop?"*) [78]. The paper exactly emphasizes the benefits of human expertise and the search for unknown patterns in a low-dimensional feature space upon which our approach is based.

We believe that it is better to use interpretable models than to explain black-box models [79]. An integration of interpretable expert systems written as optimization models with today's powerful ML approaches may result in better healthcare with interpretable results.

Conclusions

We proposed a method for the difficult classification task AFib↔AFlu that combines expert models and ML. On our gold standard test set, our approach was highly successful reaching 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 pre-classification, but cannot 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 used for determing treatments. As runtimes of the algorithm are short enough for real-time requirements, it can be applied as a decision-support tool for clinical practice. A combination of the presented feature extraction and classification with state-of-the-art NN is plausible, but open due to availability of data sets and trained models. An open question is how to further reduce failure cases due to pseudo-regularization as discussed in subsection Classification failures.

Finally, we proposed to create features from optimal solutions of domain-knowledge models and to search for unknown patterns in a low-dimensional feature space. We think 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.

Supporting information

S1 File.
(ZIP)

Author Contributions

Conceptualization: Sebastian Sager, Florian Kehrle, Eberhard Scholz.

Funding acquisition: Sebastian Sager, Eberhard Scholz.

Investigation: Sebastian Sager, Felix Bernhardt, Florian Kehrle, Eberhard Scholz.

Methodology: Sebastian Sager, Florian Kehrle, Eberhard Scholz.

Project administration: Sebastian Sager.

Resources: Sebastian Sager.

Software: Felix Bernhardt, Florian Kehrle.

Supervision: Sebastian Sager, Maximilian Merkert, Andreas Potschka, Eberhard Scholz.

Validation: Eberhard Scholz.

Writing – original draft: Sebastian Sager.

Writing – review & editing: Sebastian Sager, Florian Kehrle, Maximilian Merkert, Andreas Potschka, Benjamin Meder, Hugo Katus, Eberhard Scholz.

References

1. Mincholé A, Rodriguez B. Artificial intelligence for the electrocardiogram. Nature Medicine. 2019; 25 (1):22. https://doi.org/10.1038/s41591-018-0306-1 PMID: 30617324
2. Vaish A, Kumari P. 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. p. 1467–1476.
3. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine. 2019; 25(1):65. https://doi.org/10.1038/s41591-018-0268-3 PMID: 30617320
4. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, et al. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nature Medicine. 2019; 25(1):70. https://doi.org/10.1038/s41591-018-0240-2 PMID: 30617318
5. Fernández-Ruiz I. Artificial intelligence to improve the diagnosis of cardiovascular diseases. Nature Reviews Cardiology. 2019; p. 1. PMID: 30683888
6. Li H, Yuan D, Ma X, Cui D, Cao L. Genetic algorithm for the optimization of features and neural networks in ECG signals classification. Scientific reports. 2017; 7:41011. https://doi.org/10.1038/srep41011 PMID: 28139677
7. De Ponti R, Marazzi R, Zoli L, Caravati F, Ghiringhelli S, Salerno-Uriarte JA. Electroanatomic mapping and ablation of macroreentrant atrial tachycardia: comparison between successfully and unsuccessfully treated cases. Journal of cardiovascular electrophysiology. 2010; 21(2):155–162. https://doi.org/10.1111/j.1540-8167.2009.01602.x PMID: 19793143
8. Bumgarner JM, Lambert CT, Hussein AA, Cantillon DJ, Baranowski B, Wolski K, et al. Smartwatch algorithm for automated detection of atrial fibrillation. Journal of the American College of Cardiology. 2018; 71(21):2381–2388. https://doi.org/10.1016/j.jacc.2018.03.003 PMID: 29535065
9. Guo Y, Wang H, Zhang H, Liu T, Liang Z, Xia Y, et al. Mobile photoplethysmographic technology to detect atrial fibrillation. Journal of the American College of Cardiology. 2019; 74(19):2365–2375. https://doi.org/10.1016/j.jacc.2019.08.019 PMID: 31487545
10. Elkholey K, Lofgren MM, Meeks KQ, Asad ZUA, Freedman B, Stavrakis S. Screening for Atrial Fibrillation in Native Americans Using Smartphone-Based ECG. Circulation. 2019; 140(Suppl_1):A13895–A13895.
11. Mutke M, Brasier N, Raichle C, Doerr M, Eckstein J, research group CC. P1938 Comparing atrial fibrillation detection algorithms in smart devices on validated mobile ECG data. European Heart Journal. 2018; 39(suppl_1):ehy565–P1938. https://doi.org/10.1093/eurheartj/ehy565.P1938
12. Shiyovich A, Wolak A, Yacobovich L, Grosbard A, Katz A. 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. 2010; 340(4):271–275. https://doi.org/10.1097/MAJ.0b013e3181e73fcf PMID: 20881756

13. Knight BP, Michaud GF, Strickberger SA, Morady F. Electrocardiographic differentiation of atrial flutter from atrial fibrillation by physicians. Journal of Electrocardiology. 1999; 32:315–319. https://doi.org/10.1016/S0022-0736(99)90002-X PMID: 10549907
14. Krummen DE, Patel M, Ngyen H, Ho G, Kazi DS, Clopton P, et al. Accurate ECG diagnosis of atrial tachyarrhythmias using quantitative analysis: A prospective diagnostic and cost-effectiveness study. Journal of Cardiovascular Electrophysiology. 2010; 21:1251–1259. https://doi.org/10.1111/j.1540-8167.2010.01809.x PMID: 20522152
15. Kirchhof P, Benussi S, Kotecha D, Ahlsson A, Atar D, Casadei B, et al. 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. European Heart Journal. 2016; 37(38):2893–2962. https://doi.org/10.1093/eurheartj/ehw210 PMID: 27567408
16. Sawhney N, Anousheh R, Chen W, Feld GK. 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. 2010; 3(3):243–248.
17. Scholz EP, Kehrle F, Vossel S, Hess A, Zitron E, Katus HA, et al. Discriminating atrial flutter from atrial fibrillation using a multilevel model of atrioventricular conduction. Heart Rhythm. 2014; 11(5):877–884. https://doi.org/10.1016/j.hrthm.2014.02.013 PMID: 24561160
18. da S Luz EJ, Schwartz WR, Cámara-Chávez G, Menotti D. ECG-based heartbeat classification for arrhythmia detection: A survey. Computer Methods and Programs in Biomedicine. 2016; 127:144–164. https://doi.org/10.1016/j.cmpb.2015.12.008
19. Box GE, Cox DR. An analysis of transformations. Journal of the Royal Statistical Society: Series B (Methodological). 1964; 26(2):211–243.
20. Yeo IK, Johnson RA. A new family of power transformations to improve normality or symmetry. Biometrika. 2000; 87(4):954–959. https://doi.org/10.1093/biomet/87.4.954
21. Dougherty J, Kohavi R, Sahami M. Supervised and unsupervised discretization of continuous features. In: Machine Learning 1995 Proceedings. Elsevier; 1995. p. 194–202.
22. Garcia S, Luengo J, Sáez JA, Lopez V, Herrera F. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. IEEE Transactions on Knowledge and Data Engineering. 2012; 25(4):734–750. https://doi.org/10.1109/TKDE.2012.35
23. Estévez PA, Tesmer M, Perez CA, Zurada JM. Normalized mutual information feature selection. IEEE Transactions on Neural Networks. 2009; 20(2):189–201. https://doi.org/10.1109/TNN.2008.2005601 PMID: 19150792
24. Katz G, Shin ECR, Song D. ExploreKit: Automatic feature generation and selection. In: 2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE; 2016. p. 979–984.
25. Holzinger A. Explainable ai and multi-modal causability in medicine. i-com. 2020; 19(3):171–179. https://doi.org/10.1515/icom-2020-0024
26. Schweidtmann AM, Esche E, Fischer A, Kloft M, Repke JU, Sager S, et al. Machine Learning in Chemical Engineering: A Perspective. Chemie Ingenieur Technik. 2021. https://doi.org/10.1002/cite.202100083
27. Bikmukhametov T, Jäschke J. Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models. Computers & Chemical Engineering. 2020; 138:106834. https://doi.org/10.1016/j.compchemeng.2020.106834
28. Rackauckas C, Ma Y, Martensen J, Warner C, Zubov K, Supekar R, et al. Universal differential equations for scientific machine learning. arXiv preprint arXiv:200104385. 2020; p. 1–6.
29. Heinlein A, Klawonn A, Lanser M, Weber J. Machine Learning in Adaptive Domain Decomposition Methods—Predicting the Geometric Location of Constraints. SIAM Journal on Scientific Computing. 2019; 41(6):A3887–A3912. https://doi.org/10.1137/18M1205364
30. Raissi M, Karniadakis GE. Hidden physics models: Machine learning of nonlinear partial differential equations. Journal of Computational Physics. 2018; 357:125–141. https://doi.org/10.1016/j.jcp.2017.11.039
31. Raissi M, Perdikaris P, Karniadakis GE. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics. 2019; 378:686–707. https://doi.org/10.1016/j.jcp.2018.10.045
32. Yan S, He Y, Tang T, Wang T. Drag coefficient prediction for non-spherical particles in dense gas–solid two-phase flow using artificial neural network. Powder Technology. 2019; 354:115–124. https://doi.org/10.1016/j.powtec.2019.05.049

33. Qian E, Kramer B, Peherstorfer B, Willcox K. Lift & Learn: Physics-informed machine learning for large-scale nonlinear dynamical systems. Physica D: Nonlinear Phenomena. 2020; 406:132401. https://doi.org/10.1016/j.physd.2020.132401

34. Yazdani A, Raissi M, Karniadakis GE. Systems biology informed deep learning for inferring parameters and hidden dynamics. bioRxiv. 2019; p. 865063.

35. Kissas G, Yang Y, Hwuang E, Witschey WR, Detre JA, Perdikaris P. 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. 2020; 358:112623. https://doi.org/10.1016/j.cma.2019.112623

36. Hodgkin AL, Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. The Journal of Physiology. 1952; 117:500–544. https://doi.org/10.1113/jphysiol.1952.sp004764 PMID: 12991237

37. Villaverde AF, Banga JR. Structural Properties of Dynamic Systems Biology Models: Identifiability, Reachability, and Initial Conditions. Processes. 2017; 5(2).

38. Mazgalev T, Tchou PJ. Atrial-AV Nodal Electrophysiology: A View from the Millennium. Wiley; 2000.

39. Watanabe Y, Dreifus LS. Second degree atrioventricular block. Cardiovascular Research. 1967; 1:150–158. https://doi.org/10.1093/cvr/1.2.150 PMID: 6058851

40. Kosowsky BD, Latif P, Radoff AM. Multilevel atrioventricular block. Circulation. 1976; 54:914–921. https://doi.org/10.1161/01.CIR.54.6.914 PMID: 991406

41. Slama R, Leclercq JF, Rosengarten M, Coumel P, Bouvrain Y. Multilevel block in the atrioventricular node during atrial tachycardia and flutter alternating with Wenckebach phenomenon. British Heart Journal. 1979; 42(4):463–470. https://doi.org/10.1136/hrt.42.4.463 PMID: 508477

42. Littmann L, Svenson RH. Atrioventricular alternating Wenckebach periodicity: Conduction patterns in multilevel block. The American Journal of Cardiology. 1982; 49(4):855–862. https://doi.org/10.1016/0002-9149(82)91969-5 PMID: 7064834

43. Castellanos A, Diaz J, Interian A, Myerburg RJ. Wenckebach's periods or alternating Wenckebach's periods during 4:1 atrioventricular block? Journal of Electrocardiology. 2005; 38:157–159. https://doi.org/10.1016/j.jelectrocard.2004.10.007 PMID: 15892027

44. Wolkenhauer O, Auffray C, Brass O, Clairambault J, Deutsch A, Drasdo D, et al. Enabling multiscale modeling in systems medicine. Genome medicine. 2014; 6(3):21. https://doi.org/10.1186/gm538 PMID: 25031615

45. Fröhlich F, Theis FJ, Rädler JO, Hasenauer J. Parameter estimation for dynamical systems with discrete events and logical operations. Bioinformatics. 2017; 33(7):1049–1056. PMID: 28040696

46. Kremling A, Geiselmann J, Ropers D, de Jong H. An ensemble of mathematical models showing diauxic growth behaviour. BMC systems biology. 2018; 12(1):1–16. https://doi.org/10.1186/s12918-018-0604-8 PMID: 30241537

47. Tsipa A, Pitt JA, Banga JR, Mantalaris A. 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. 2020;. https://doi.org/10.1007/s00449-020-02360-2 PMID: 32377941

48. Josephson ME. Clinical Cardiac Electrophysiology: Techniques and Interpretations. 4th ed. Lippincott Williams & Wilkins; 2008.

49. Wenckebach KF, Winterberg H. Die unregelmäßige Herztätigkeit. Wilhelm Engelnmann; 1927.

50. Denes P, Levy L, Pick A, Rosen KM. The incidence of typical and atypical A-V Wenckebach periodicity. American Heart Journal. 1975; 89(1):26–31. https://doi.org/10.1016/0002-8703(75)90005-8 PMID: 1109548

51. Spodick DH. Seven-Cycle Wenckebach Period Without Atypical Features. American Heart Hospital Journal. 2004; 2(1):64. https://doi.org/10.1111/j.1541-9215.2004.03394.x PMID: 15604845

52. Friedman HS, Gomes JAC, Haft JI. An Analysis of Wenckebach Periodicity. Journal of Electrocardiology. 1975; 8(4):307–315. https://doi.org/10.1016/S0022-0736(75)80003-3 PMID: 1176840

53. Hay J. Bradycardia and cardiac arrhythmias produced by depression of certain functions of the heart. Lancet. 1906; 1:138–143.

54. Barold SS, Lüderitz B. John Hay and the Earliest Description of Type II Second-Degree Atrioventricular Block. The American Journal of Cardiology. 2001; 87(12):1433–1435. https://doi.org/10.1016/S0002-9149(01)01574-0 PMID: 11397375

55. de Medina EOR, Bernard R, Coumel P, Damato AN, Fisch C, Krikler D, et al. WHO/ISC Task Force. Definition of terms related to cardiac rhythm. American Heart Journal. 1978; 95:796–806. https://doi.org/10.1016/0002-8703(78)90512-4

56. Surawicz B, Uhley H, Borun R, Laks M, Crevasse L, Rosen K, et al. The quest for optimal standardization of terminology and interpretation. American Heart Journal. 1978; 41(1):130–145. PMID: 622995

57. Zipes DP, Dimarco JP, Gillette PC, Jackman WM, Myerburg RJ, Rahimtoola SH, et al. Guidelines for clinical intracardiac electrophysiological and catheter ablation procedures. Journal of the American College of Cardiology. 1995; 26(2):555–573. https://doi.org/10.1016/0735-1097(95)80037-H PMID: 7608464

58. Barold SS. 2:1 Atrioventricular Block: Order from Chaos. The American Journal of Emergency Medicine. 2001; 19(3):214–217. https://doi.org/10.1053/ajem.2001.21715 PMID: 11326349

59. Kehrle F. Inverse Simulation for Cardiac Arrhythmia. Otto-von-Guericke University Magdeburg; 2018. Available from: https://mathopt.de/PUBLICATIONS/Kehrle2018.pdf.

60. Sager S; Mathematical Optimisation Society. Optimization and Clinical Decision Support. Optima. 2018; 104:1–8.

61. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011; 2:27:1–27:27. https://doi.org/10.1145/1961189.1961199

62. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995; 20(3):273–297. https://doi.org/10.1007/BF00994018

63. Camm AJ, Kirchhof P, G YHL et al. Guidelines for the management of atrial fibrillation. European Heart Journal. 2010; 31:2369–2429. https://doi.org/10.1093/eurheartj/ehq278 PMID: 20802247

64. Goldberger A, Amaral L, Glass L, Hausdorff J, Ivanov PC, Mark R, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 2000; 101(23):e215–e220. PMID: 10851218

65. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. CoRR. 2014;abs/1412.6980.

66. Hoppe BL, Kahn AM, Feld GK, Hassankhani A, Narayan SM. Separating atrial flutter from atrial fibrillation with apparent electrocardiographic organization using dominant and narrow F-wave spectra. Journal of the American College of Cardiology. 2005; 46:2079–2087. https://doi.org/10.1016/j.jacc.2005.08.048 PMID: 16325046

67. Bogun F, Anh D, Kalahasty G, Wissner E, Serhal CB, Bazzi R, et al. Misdiagnosis of atrial fibrillation and its clinical consequences. The American Journal of the Medical Sciences. 2004; 117(9):636–642. PMID: 15501200

68. Kettering K, Dörnberger V, Lang R, Vonthein R, Suchalla R, Bosch RF, et al. Enhanced detection criteria in implantable cardioverter defibrillators: Sensitivity and specificity of the stability algorithm at different heart rates. Pacing and Clinical Electrophysiology. 2001; 24:1325–1333. https://doi.org/10.1046/j.1460-9592.2001.01325.x PMID: 11584454

69. Ahmed S, Claughton A, Gould PA. Atrial flutter—diagnosis, management and treatment. In: Abnormal Heart Rhythms. IntechOpen; 2015.

70. Garcia-Cosio F, Fuentes AP, Angulo AN. Clinical approach to atrial tachycardia and atrial flutter from an understanding of the mechanisms. Electrophysiology based on anatomy. Revista Española de Cardiologia (English Edition). 2012; 65(4):363–375. https://doi.org/10.1016/j.rec.2011.11.013 PMID: 22364957

71. Hornik K. Approximation capabilities of multilayer feedforward networks. Neural networks. 1991; 4 (2):251–257. https://doi.org/10.1016/0893-6080(91)90009-T

72. Kidger P, Lyons T. Universal approximation with deep narrow networks. In: Conference on Learning Theory; 2020. p. 2306–2327.

73. Jagtap AD, Kawaguchi K, Karniadakis GE. Adaptive activation functions accelerate convergence in deep and physics-informed neural networks. Journal of Computational Physics. 2020; 404:109136. https://doi.org/10.1098/rspa.2020.0334 PMID: 32831616

74. Jost F, Zierk J, Le TTT, Raupach T, Zierk J, Rauh M, et al. Model-based simulation of maintenance therapy of childhood acute lymphoblastic leukemia. Frontiers in Physiology. 2020; 11:217. https://doi.org/10.3389/fphys.2020.00217 PMID: 32256384

75. Jost F, Schalk E, Weber D, Döhner H, Fischer T, Sager S. Model-based optimal AML consolidation treatment. IEEE Transactions on Biomedical Engineering. 2020; 67:3296–3306. https://doi.org/10.1109/TBME.2020.2982749 PMID: 32406820

76. Lilienthal P, Tetschke M, Schalk E, Fischer T, Sager S. Optimized and Personalized Phlebotomy Schedules for Patients suffering from Polycythemia Vera. Frontiers in Physiology. 2020; 11:328. https://doi.org/10.3389/fphys.2020.00328 PMID: 32362837

77. Littmann M, Selig K, Cohen-Lavi L, Frank Y, Hönigschmid P, Kataka E, et al. Validity of machine learning in biology and medicine increased through collaborations across fields of expertise. Nature Machine Intelligence. 2020; 2(1):18–24. https://doi.org/10.1038/s42256-019-0139-8

78. Holzinger A. Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Informatics. 2016; 3(2):119–131. https://doi.org/10.1007/s40708-016-0042-6 PMID: 27747607

79. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019; 1:206–215. https://doi.org/10.1038/s42256-019-0048-x