Expert-Enhanced Machine Learning for Cardiac Arrhythmia Classification

Sebastian Sager, Felix Bernhardt, Florian Kehrle, Maximilian Merkert, Andreas Potschka, Benjamin Meder, Hugo Katus, and Eberhard Scholz

Abstract—Machine learning (ML) methodology has been successfully applied to many classification problems in medicine and beyond. Whereas the accuracy is often astonishing, the interpretability of the results has become an ubiquitous issue. In order to overcome this important but unsolved challenge, we propose to first reduce the complexity of the data, and then to combine the interpretability of expert systems with the deductive power of data driven ML. As a showcase we considered the arguably most difficult classification case of cardiac arrhythmias. Here the largest database with the gold standard (intracardiac measurements after invasive procedures) only contains 380 samples, yielding an additional challenge to ML. Still, our approach achieved an accuracy of 82.84%. The main advantage however is the interpretability of the classification results. Our features give insight into a possibly occurring multi-level atrioventricular blocking mechanism, which might improve treatment decisions and is thus an important step in the realization of personalized medicine. The idea to use mathematical modeling and optimization to generate new and clinically interpretable features for ML can be transferred to other cases of clinical decision support.

Index Terms—Optimization, Machine learning, Healthcare, Decision support, Combinatorial algorithms.

I. INTRODUCTION

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]–[4] 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. Parameters like the atrial cycle length were not provided, although they might be relevant for treatment decisions [5].

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 (AFlu) and focal atrial tachycardia with irregular ventricular response. Usually both physiological cases are lumped together both in deep learning (DL), “The atrial fibrillation class combined atrial fibrillation and atrial flutter” [2], and in heart rate variability based algorithms for smartwatches [6]. 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, compare Figure 1. High misdiagnosis rates and possible causes have been reported [7]–[9]. This is concerning, as different treatments (often antiarrhythmics in AFib and ablation in AFlu) are implied by the diagnosis [10] and atypical forms of AFlu are becoming increasingly important in clinical practice as a complication of left atrial ablation procedures [11].

The poor quality of expert opinion due to the difficult discrimination poses also a challenge to automated classification by supervised ML, which often uses it for labeling training samples [2]–[4]. We used an expert analysis based on intracardiac measurements as gold standard, which is only available with invasive procedures.

It was shown that DL can robustly distinguish samples that are either AFib or AFlu from other cardiac arrhythmias [2]. We propose to extend and complement this approach with generated features that are based on a pathophysiological rationale allowing to also classify AFib ↔ AFlu. In the following we are going to assume that it has already been verified that either AFib or AFlu is present. The most important building block 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” [12]. Key to solving this riddle is the idea of a multi-level AV block (MAVB) [13]–[17]. The tedious procedure of manually adjusting possible MAVB combinations has been automatized with large success in the algorithm HEAT [18]–[20]. The combination of mathematical model and numerical optimization algorithm could also be seen as an interpretable expert system. Here, for the first time we used HEAT with an improved mathematical model of MAVB for ML feature generation. 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 [13]–[17]. We considered optimization variables...
Fig. 1. 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), 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 $\Delta a$, the cycle length in the atrial chambers (A), is regular and the forward simulation in V is close to rawRR (minimal deviation $F(x)$). 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), compare bottom zooms. It can be visually confirmed that for AFib the calculated $\Delta a$ corresponds well to the intracardiac measurements.

$x = (\Delta a, bt, oc)$ with $\Delta a$ atrial cycle length, $bt$ blocktype, and $oc$ a vector of blocktype-specific internal offset counters and conduction constants. For different choices of $x$ a forward simulation of ventricular responses (RR interval lengths) is possible, which can be compared to the RR measurements, see Figure 1. A penalization of the difference in an appropriate metric like the Euclidean norm gives an objective function $F(x)$. In an inverse simulation HEAT calculated for all training samples $i$ optimal solutions $x_i^*$, i.e., feasible values $\Delta a_i^*, bt_i^*$, and $oc_i^*$ which resulted in the smallest deviations $F_i(x_i^*)$.

II. METHODS

A. Data

Our data set heatDS is a superset of the one used in a previous study [18], 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 160 patients.

For all 160 patients the classification AFib$\leftrightarrow$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 ($\Delta a$) 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 [21]. 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 $\Delta a$ 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.

B. Multilevel Atrioventricular Block (MAVB)

We developed a mathematical model for MAVB, which is 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 cannot 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 [22]. 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.

- This theoretical concept allows to combine different blocking ratios \( n+1:n \) (possibly varying and with linearly changing conduction times due to RRP) on an unlimited number of levels. However, it makes sense to limit the number of possible combinations to avoid overfitting, to reduce computational times, and to stay close to clinical observations. We restricted our MAVB model to the five combinations shown in Figure 2 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 [23]–[25], atypical Type I block [24], [26], the special cases of 2:1 and 3:2 Type I blocks, Type II block [27]–[30], advanced second-degree AV Block [31]–[33], and MAVB [13]–[17]. Preferable in the sense of Ockham’s razor, this unified model also allows an efficient calculation of the most likely block for given RR data.

C. HEAT

In an inverse simulation HEAT calculated for all training samples \( i \) the optimal solution \( x^*_i \), i.e., the particular values \( \Delta a^*_i, bt^*_i \), and \( oc^*_i \) which resulted in the smallest objective function value

\[
F_i(x^*_i) = \min_{x \in \mathcal{X}} F_i(x).
\]

Here \( \mathcal{X} \) is 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 2. The bounds on the atrial cycle length \( \Delta a \) were determined based on physiological observations [22] (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 \( \Delta a \) and \( oc \). The algorithm HEAT is patented [19].

D. Features and Feature Sets

We investigated the following features

- The time series of raw input RR interval times (RR), together with the derived scalar features heart rate variability (RRvar) and average heart rate (RMean).
- HEAT optimal objective function value \( F(x^*) \) (HEATobj)
- 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 \( n_{RR} = 22 \) time intervals, we considered only \( n_{sub} \in I := \{10, \ldots, n_{RR} \} \) 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 \( i \in I \). To investigate the robustness of solutions, we also evaluated \( F_{i,j,k}(x^*_{i,k}) \) for \( j, k \in I \), 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 \( i^{\prime} \)” (HEATfit)). 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 considered age (age).

In summary, we implemented the following sets of features

- rawRR = [RR]
- heatObjective = [HEATobj]
- heatSolution = [HEATobj, HEATsol, RRvar, RMean]
- heatSerAvg = [HEATseriesAvg]
- heatSerAvgAge = [HEATseriesAvg, age]
- heatSeries = [HEATseries]
TABLE I
AVERAGE ACCURACIES, STANDARD DEVIATIONS, AND NUMBERS OF OPTIMIZATION VARIABLES

| Feature Set    | ML Model | Accuracy   | Sensitivity | Specificity | Parameters | Scaling | Hyperpars |
|----------------|----------|------------|-------------|-------------|------------|---------|-----------|
| rawRR          | CNN      | 57.26% ± 6.47% | 51.68% ± 29.04% | 62.84% ± 27.25% | 287–487    | 0       | 2         |
| rawRR          | SVM N-Gram | 62.03% ± 5.25% | 70.52% ± 11.83% | 53.53% ± 15.49% | 101–485    | 200–968 | 4         |
| heatObjective  | SVM      | 77.58% ± 4.15% | 80.26% ± 6.87% | 74.89% ± 7.33% | 2          | 2       | 4         |
| heatSolution   | SVM      | 79.37% ± 4.55% | 85.05% ± 7.03% | 73.68% ± 6.96% | 10         | 18      | 4         |
| heatSerAvg     | SVM      | 82.18% ± 4.48% | 88.16% ± 5.77% | 76.21% ± 6.40% | 21         | 40      | 4         |
| heatSerAvgAge  | SVM      | 82.47% ± 3.26% | 87.84% ± 5.18% | 77.10% ± 5.70% | 23         | 44      | 4         |
| heatSeries     | SVM N-Gram | 82.84% ± 4.31% | 87.21% ± 6.09% | 78.47% ± 7.69% | 91–1691    | 180–3380 | 4         |

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

- SVM: 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 $\geq 1$ and $\leq n_{RR}$ was considered. The standard deviation was only computed on subsequences of length $\geq 2$. 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 [34]. The underlying model is described in [35]. The kernel type (radial basis functions or polynomial) with a penalty parameter $C$ and a kernel coefficient $\gamma$ (3 values each) and the length of analyzed subsequences $n_{sub} \in \{10, \ldots, 22\}$ were tuned as hyperparameters using grid search cross-validation.

- CNN: We used a CNN architecture consisting of a sequence 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 I 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.

F. 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 CNN were based on tensorflow 1.8.0 and trained using the Adam optimizer [36] with default parameters.

The computational times were roughly 1 second per HEAT call (times 380 samples times number of considered subproblems per sample), 30 minutes for training SVM, and 3 days for training CNN.

III. RESULTS
We show the mean sensitivities and specificities (and the resulting averaged accuracies) in Table I. The results were obtained after repeated, stratified 10-fold cross validation for different feature sets and ML models. Features were time series of raw input RR interval times (RR), together with the derived scalar features heart rate variability (RRvar) and average heart rate (RRmean); HEAT optimal objective function value $F(x^*)$ (HEATobj); HEAT optimal solution $x^* = (\Delta a^*, bt^*, oc^*)$ (HEATsol), a series of HEAT solutions on smaller time horizons (HEATseries), and averaged values (HEATseriesAvg). Feature sets were rawRR = [RR], as well as heatObjective = [HEATobj], heatSolution = [HEATobj, HEATsol, RRvar, RRmean], heatSerAvg = [HEATseriesAvg], heatSerAvgAge = [HEATseriesAvg, age], and heatSeries = [HEATseries]. Different established ML models were applied, most importantly a Support Vector Classifier (SVM), a variant for time series (SVM N-Gram), and a Convolutional Neural Network (CNN).

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 $F_1(x^*)$ was used as the only feature (generated a priori from rawRR). A higherdimensional classification, which also took $x^*$ 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 3.

For an implementation of a convolutional neural network (CNN) the poor performance of direct application to rawRR...
Fig. 3. Representative pairwise plot of features obtained from a heatSolution SVM classification, compare Table I. One observes the clear separation of atrial flutter (AFlu) and atrial fibrillation (AFib) with respect to the feature HEAT optimal objective function value $F(x^*)$ (HEATobj). The two model parameters in $x^*$, the atrial cycle length $\Delta a$ and the blocktype $bt$ do not allow a straightforward classification. AFib is often misdiagnosed with one particular out of five possible blocktypes $bt$ (the two-level block MAV2A). This is due to a larger range of possible time series that can be simulated with this blocktype and one possible target to reduce misclassification of AFib.

was also reflected by high standard deviations. The number of ML parameters was two orders of magnitude larger 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 [2]. The approach to preprocess $\text{rawRR}$ using medical expert knowledge (HEAT) can thus also be seen as an approach to increase sensitivity without overfitting the ML model.

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 $\Delta a^*$ might be helpful for an a posteriori identification of flutter waves for AFlu in a surface ECG. The optimal blocktype $bt^*$, compare Figures 4 and 5 with two and three levels with varying blockings, respectively, gives insight into the pathophysiology of the AV node.

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

IV. DISCUSSION

A. 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 [10].

All ML approaches that were applied directly to the input data ($\text{rawRR}$) resulted in average accuracies of approximately 60%. These low accuracies were not surprising, as AFib↔AFlu is a difficult case even for experts [7]–[9] and was explicitly excluded in recent studies [2]. AFib may be overdiagnosed because of coarse fibrillatory waves which are reminiscent of AFlu [8], [37], the presence of artifacts, or premature atrial complexes [38]. AFlu may be overdiagnosed, because the low-voltage flutter waves that indicate AFib can be hardly discernible in the surface ECG, compare Figures 1 and 4. The achieved accuracies are similar to previous results to analyze AFib↔AFlu, e.g., based on clustering of RR times or nodal recovery approaches [20]. Note that the N-Gram approach implicitly considers $\text{RRvar}$, $\text{RRmean}$ and is thus
Fig. 4. As in Figure 1 left, but for different input data from the same patient. The 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 $\Delta a$ proposed by HEAT (leftmost zoom). When no intracardiac measurements are available, this value $\Delta a$ could be of help for the physician, e.g., when carefully reanalyzing the ECG. An overlay of $\Delta a$ makes the task to spot atrial activations in the surface ECG easier (middle zoom).

a superset of features used in current smartwatch algorithms [6]. Hence, the low accuracy gives a hint why AFib$\Leftrightarrow$AFlu cannot be treated by them. To summarize, for short RR time series and comparably few labeled training samples none of the mentioned approaches seems to be capable of providing a high classification accuracy for AFib$\Leftrightarrow$AFlu.

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 $\Delta a$ 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. A limit for the classification accuracy arises from false positives, e.g., 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.

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 a few seconds 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 [2]. It can use the prior classification of AFib and AFlu into one cluster, and can classify AFib$\Leftrightarrow$AFlu in a following step. HEAT can be run on a server. A secure client-server architecture has been implemented [20]. 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 3, 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.

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
Fig. 5. As in Figures 1 left and 4, 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). Again, the intracardiac measurements are shown for illustrative purposes (top). The close match to the calculated atrial cycle length $\Delta a$ highlights the plausibility of the complex blocking mechanism. The calculated blocktype might give insight into the pathophysiology of the AV node and be useful for choosing a good treatment.

were large enough, but even then the approach would lack interpretability.

B. 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 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 $\Delta a^*$ 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 [5]. In addition, for AFib “a thorough understanding of electrophysiological properties and anatomical landmarks is essential in achieving a successful ablation outcome and in reducing complication rates” [39]. Sometimes it is even claimed that “the classic ECG-based diagnoses of tachycardia and AFib are of little importance today because treatment is based on the direct management of the trigger mechanism” [40]. 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.

C. Generalization to other cases of clinical decision support

Led by text classification and image processing, ML has been conquering many areas of modern life and the sciences. Despite some disappointments [41], 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 (AI) in modern life and research. 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 [42], in particular in healthcare applications [43], [44]. 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, 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.” [45]. Similar statements can also be found in the German national AI strategy report [46].

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 arrhythmia could be done in a similar way as above. But also for diseases like acute leukemias or polycythemia vera 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 [47].

We believe that it is better to use interpretable models than to explain black box models [48]. An integration of interpretable expert systems written as optimization models with today’s powerful ML approaches might result in better healthcare with interpretable results.

ACKNOWLEDGMENT

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

REFERENCES

[1] A. Mincholé and B. Rodriguez, “Artificial intelligence for the electrocardiogram,” *Nature Medicine*, vol. 25, no. 1, p. 22, 2019.
[2] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, “Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network,” *Nature Medicine*, vol. 25, no. 1, p. 65, 2019.
[3] Z. I. Atta, 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*, vol. 25, no. 1, p. 70, 2019.
[4] I. Fernández-Ruíz, “Artificial intelligence to improve the diagnosis of cardiovascular diseases,” *Nature Reviews Cardiology*, p. 1, 2019.
[5] R. De Ponti, R. Marazzi, L. Zoli, F. Caravati, S. Ghiringhelli, and J. A. Salerno-Urriate, “Electroanatomic mapping and ablation of macroreentrant atrial tachycardia: comparison between successful and unsuccessfully treated cases,” *Journal of Cardiovascular Electrophysiology*, vol. 21, no. 2, pp. 155–162, 2010.
[6] J. M. Bungunger, C. T. Lambert, A. A. Hussein, D. J. Cantillon, B. Baranowski, K. Wolski, B. D. Lindsay, O. M. Wazni, and K. G. Tarakji, “Smartwatch algorithm for automated detection of atrial fibrillation,” *Journal of the American College of Cardiology*, vol. 71, pp. 2381–2388, 2018.
[7] A. Shiyovich, A. Wolak, L. Yacobovich, A. Grosbard, and 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*, vol. 340, no. 4, pp. 271–275, 2010.
[8] B. Knight, G. Michaud, S. Strickberger, and F. Morady, “Electrocardiographic differentiation of atrial flutter from atrial fibrillation by physicians,” *Journal of Electrocardiology*, vol. 32, pp. 315–319, 1999.
[9] D. Krummen, M. Patel, H. Ngyen, G. Ho, D. Kazi, P. Clifton, M. Holland, S. Greenberg, G. Feld, M. Faddis, and S. Narayan, “Accurate ECG diagnosis of atrial tachyarrhythmias using quantitative analysis: A prospective diagnostic and cost-effectiveness study,” *Journal of Cardiovascular Electrophysiology*, vol. 21, pp. 1251–1259, 2010.
[10] 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*, vol. 37, no. 38, pp. 2893–2962, 08 2016. [Online]. Available: https://doi.org/10.1093/eurheartj/ehw210
[11] N. Sawhney, R. Anousheh, W. Chen, and 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*, vol. 3, no. 3, pp. 243–248, 2010.
[12] T. Mrazek and P. J. Towne, *Atrial-VNodal Electrophysiology: A View from the Millennium*. Wiley, 2000.
[13] Y. Watanabe and L. Dreifus, “Second degree atrioventricular block,” *Cardiovascular Research*, vol. 1, pp. 150–158, 1967.
[14] B. Kosowsky, P. Latif, and A. Radoff, “Multilevel atrioventricular block,” *Circulation*, vol. 54, p. 914–921, 1976.
[15] R. Slama, J. F. Leclercq, M. Rosengarten, P. Coumel, and Y. Bouvrand, “Multilevel block in the atrioventricular node during atrial tachycardia and flutter alternating with wenckebach phenomenon,” *British Heart Journal*, vol. 42, no. 4, pp. 463–470, 1979.
[16] L. Littmann and R. H. Svenson, “Atrioventricular alternating Wenckebach periodicity: Conduction patterns in multilevel block,” *The American Journal of Cardiology*, vol. 49, no. 4, pp. 855–862, 1982.
[17] A. Castellanos, J. Diaz, A. Interian, and R. J. Myerburg, “Wenckebach’s periods or alternating wenckebach’s periods during 4:1 atrioventricular block?” *Journal of Electrophysiology*, vol. 15, pp. 157–159, 1994.
[18] E. Scholz, F. Kehrle, S. Vossal, A. Hess, E. Zitron, H. Katus, and S. Sager, “Discriminating atrial flutter from atrial fibrillation using a multilevel model of atrioventricular conduction,” *Heart Rhythm*, vol. 11, no. 5, pp. 877–884, 2014. [Online]. Available: https://matshop.de/PUBLICATIONS/Scho14_2014.pdf
[19] E. Scholz, S. Sager, and H. Katus, “A system and computer program product for automatically distinguishing atrial flutter from atrial fibrillation,” European Patent EP2757940B1, 2016. [Online]. Available: https://patents.google.com/patent/EP2757940B1
[20] F. Kehrle, “Inverse simulation for cardiac arrhythmia,” Ph.D. dissertation, Otto-von-Guericke University Magdeburg, 2018. [Online]. Available: https://matshop.de/PUBLICATIONS/Kehrle2018.pdf
[21] A. Camm, P. Kirchhof, and G. L. et al., “Guidelines for the management of atrial fibrillation,” *European Heart Journal*, vol. 31, pp. 2369–2429, 2010.
[22] M. E. Josephson, *Clinical Cardiac Electrophysiology: Techniques and Interpretations*, 4th ed. Lippincott Williams & Wilkins, 2008.
[23] K. F. Wenckebach and H. Winterberg, *Die unregelmäßige Herzrhythmik*. Wilhelm Engelmann, 1927.
[24] P. Denes, L. Levy, A. Pick, and K. M. Rosen, “The incidence of typical and atypical A-V Wenckebach periodicity,” *American Heart Journal*, vol. 89, no. 1, pp. 26–31, 1975.
[25] D. H. Spodick, “Seven-cycle wenckebach period without atypical features,” *American Heart Hospital Journal*, vol. 2, no. 1, p. 64, 2004.
[26] H. S. Friedman, J. A. C. Gomes, and J. I. Haft, “An analysis of Wenckebach periodicity,” *Journal of Electrocardiology*, vol. 8, no. 4, pp. 307–315, 1975.
[27] J. Hay, “Bradycardia and cardiac arrhythmias produced by depression of certain functions of the heart,” *Nature*, vol. 138–143, 1906.
[28] S. S. Barold and B. L. À¿jerzité, “John hay and the Earliest Description of Type II Second-Degree Atrioventricular Block,” *The American Journal of Cardiology*, vol. 87, no. 12, pp. 1433–1435, 2001.
[29] 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, and H. Wellens, “Who’s task force: Definition of terms related to cardiac rhythm,” *American Heart Journal*, vol. 95, pp. 796–806, 1978.
[30] B. Sauwicz, H. Uhley, R. Borun, M. Laks, L. Crevasse, K. Rosen, W. Nelson, W. Mandel, P. Lawrence, L. Jackson, N. Flowers, J. Clifton, J. Greenfield, and E. R. D. Medina, “The quest for optimal standardization of terminology and interpretation,” *American Heart Journal*, vol. 41, no. 1, pp. 130–145, 1978.
[33] S. S. Barold. “2:1 Atrioventricular block: Order from chaos,” The American Journal of Emergency Medicine, vol. 19, no. 3, pp. 214–217, 2001.

[34] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” ACM Transactions on Intelligent Systems and Technology, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[35] C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 9 1995. [Online]. Available: https://doi.org/10.1007/BF00994018

[36] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” CoRR, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[37] B. L. Hoppe, A. Kahn, G. Feld, A. Hassankhani, and 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, vol. 46, pp. 2079–2087, 2005.

[38] F. Bogun, D. Anh, G. Kalahasty, E. Wissner, C. B. Serhal, R. Bazzi, W. Weaver, and C. Schuger, “Misdiagnosis of atrial fibrillation and its clinical consequences,” The American Journal of the Medical Sciences, vol. 117, no. 9, pp. 636–642, 2004.

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

[40] F. Garcia-Cosio, A. P. Fuentes, and A. N. Angulo, “Clinical approach to atrial tachycardia and atrial flutter from an understanding of the mechanisms. electrophysiology based on anatomy,” Revista Española de Cardiología (English Edition), vol. 65, no. 4, pp. 363–375, 2012.

[41] E. Strickland, “Ibm watson, heal thyself: How ibm overpromised and underdelivered on ai health care,” IEEE Spectrum, vol. 56, no. 04, pp. 24–31, 2019.

[42] 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,” USDOE Office of Science (SC), Washington, DC (United States), Tech. Rep., 2019.

[43] E. J. Topol, “High-performance medicine: the convergence of human and artificial intelligence,” Nature Medicine, vol. 25, no. 1, pp. 44–56, 2019. [Online]. Available: https://doi.org/10.1038/s41591-018-0300-7

[44] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang, “The practical implementation of artificial intelligence technologies in medicine,” Nature Medicine, vol. 25, no. 1, pp. 30–36, 2019. [Online]. Available: https://doi.org/10.1038/s41591-018-0307-0

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

[46] Bundesregierung der Bundesrepublik Deutschland. (2018) Nationale strategie färäij kääinjstliche intelligentz. [Online]. Available: https://www.ki-strategie-deutschland.de/home.html

[47] S. Sager, “Optimization and clinical decision support,” Optima, vol. 104, pp. 1–8, 2018. [Online]. Available: http://www.mathopt.org/Optima-Issues/optima104.pdf

[48] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” Nature Machine Intelligence, vol. 1, pp. 206–215, 2019.