Meeting the unmet needs of clinicians from AI systems showcased for cardiology with deep-learning–based ECG analysis

Yonatan Elul[a,1], Aviv A. Rosenberg[a,1], Assaf Schuster[a], Alex M. Bronstein[a], and Yael Yaniv[b,2]

[a]Computer Science, Technion – Israel Institute of Technology, Haifa, 3200003, Israel; and [b]Biomedical Engineering, Technion – Israel Institute of Technology, Haifa, 3200003, Israel

Edited by David L. Donoho, Stanford University, Stanford, CA, and approved April 7, 2021 (received for review October 2, 2020)

Despite their great promise, artificial intelligence (AI) systems have yet to become ubiquitous in the daily practice of medicine largely due to several crucial unmet needs of healthcare practitioners. These include lack of explanations in clinically meaningful terms, handling the presence of unknown medical conditions, and transparency regarding the system's limitations, both in terms of statistical performance as well as recognizing situations for which the system's predictions are irrelevant. We articulate these unmet clinical needs as machine-learning (ML) problems and systematically address them with cutting-edge ML techniques. We focus on electrocardiogram (ECG) analysis as an example domain in which AI has great potential and tackle two challenging tasks: the detection of a heterogeneous mix of known and unknown arrhythmias from ECG and the identification of underlying cardio-pathology from segments annotated as normal sinus rhythm recorded in patients with an intermittent arrhythmia. We validate our methods by simulating a screening for arrhythmias in a large-scale population while adhering to statistical significance requirements. Specifically, our system 1) visualizes the relative importance of each part of an ECG segment for the final model decision; 2) upholds specified statistical constraints on its out-of-sample performance and provides uncertainty estimation for its predictions; 3) handles inputs containing unknown rhythm types; and 4) handles data from unseen patients while also flagging cases in which the model's outputs are not usable for a specific patient. This work represents a significant step toward overcoming the limitations currently impeding the integration of AI into clinical practice in cardiology and medicine in general.

artificial intelligence | medical | cardiology | deep learning

For decades, researchers have been applying machine-learning (ML) algorithms to medical tasks with the goal of incorporating insights derived from data into real-world medical applications (1). Medical artificial intelligence (AI) can potentially be used to increase personalization, reduce physician cognitive load, aid decision-making, enable preventive medicine through predictions, automate analysis of medical images and health records, and much more (2–5). Recently, deep learning (DL), a branch of ML focusing on algorithms that can learn directly from raw data (e.g., physiological signals), has risen to prominence. Such algorithms are responsible for many of the current state-of-the-art results reported in the literature for medical AI tasks, with several works claiming to surpass a human doctor's performance (5) in cases such as automatic diagnosis of breast cancer from mammography scans (6); of melanoma from skin images (7); of pathology from optical coherence tomography scans (8); assessing rehospitalization and in-hospital death risks via analysis of electronic health records (9); arrhythmia detection from electrocardiograms (ECG) analysis (10); and more. However, this slew of research successes has so far not led to widespread adoption of DL-based solutions in the day-to-day practice of medicine and in the healthcare industry in general. Over the past two years, the clinical community has responded to the recent results reported by AI researchers via several top-tier review and opinion papers (1, 2, 11–13), in which leading clinicians addressed some specific shortcomings in recent state-of-the-art medical AI solutions, which prevent them from being incorporated into clinical practice.

Here, we consolidate the most frequently mentioned shortcomings voiced by leading physicians and researchers into what we term the "unmet needs" of clinicians from medical AI. We focus specifically on shortcomings that most critically prevent clinical adoption of medical AI solutions according to the recent medical literature. We formulate these unmet needs in a manner actionable by ML researchers. We define each unmet need accurately, demonstrate how it can be met for electrocardiogram (ECG) analysis tasks, and describe our contributions with respect to it. Our work goes beyond the usual search for better model architectures or improved accuracy and focus on the challenges of clinical usefulness. To clarify, although widespread adoption of AI in the clinic is also impeded by broad systemic issues such as regulatory, legal or ethical considerations, lack of standardized data formats, integration with existing infrastructure, and others, this work focuses only on the shortcomings of the AI systems themselves.

Clinical Interpretability: Explaining Model Predictions with Medical Notions and Terminology

Arguably the loudest-voiced concern, especially regarding DL-based solutions, is the lack of model interpretability, meaning there is no way for the clinician to understand which factors led to the AI

Significance

The use of artificial intelligence (AI) in medicine, particularly deep learning, has gained considerable attention recently. Although some works boast superior capabilities compared to clinicians, actual deployments of AI systems in the clinic are scarce. We describe four important gaps on the machine-learning side responsible for this discrepancy by first formulating them in a way that is actionable by AI researchers and then systematically addressing these needs. Aiming beyond the search for better model architectures or improved accuracy, we focus directly on the challenges of clinical usefulness as stated by medical professionals in the literature. Our results show that deep-learning systems can be robust, trustworthy, explainable, and transparent while retaining the superior level of performance these algorithms are known for.

Author contributions: Y.E., A.S., A.M.B., and Y.Y. designed research; Y.E. and A.A.R. performed research; Y.E. and A.A.R. analyzed data; and Y.E., A.A.R., and Y.Y. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

See online for related content such as Commentaries.

[1]Y.E. and A.A.R. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: yaely@bm.technion.ac.il.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2020620118/-/DCSupplemental.

Published June 7, 2021.

system's decision (2, 14). This issue becomes doubly relevant in light of the European Union's General Data Protection Regulation from April 2016, the 22nd article of which stipulates that a subject of data-driven algorithmic decisions has a right to both human intervention and meaningful explanation regarding said decisions (15). Unfortunately, interpretability is an elusive concept; as yet, the field of ML holds no consensus regarding its definition (16). Furthermore, clinicians' views regarding this matter are mainly about justifying a model's output using medical-domain knowledge familiar to them (14). While common opinion holds that DL algorithms are uninterpretable "black-boxes" (11, 17, 18), this depends on what is required to be explained, with no lack of evidence that deep models are, in some respects, more interpretable than traditional models (16, 19). Here, we define this need as providing a clinician with a clear relationship between 1) the human-recognizable features of the input, 2) the relevant medical domain knowledge, and 3) the model's output. For example, an automated diagnosis of an arrhythmia (3) should be accompanied by the ECG morphology in the input data most contributing to that diagnosis (1), which should ideally be an established symptom for the detected arrhythmia in the literature (2).

Uncertainty Estimation

The next often-cited issue is that of model uncertainty estimation: a model's ability to specify how certain it is of any prediction it generates. ML models can output probabilities, which are often interpreted as the model's confidence level for each class (e.g., medical condition). Ideally, these probabilities would be calibrated: given 10 samples for which a model specified a probability of 0.8 for class c, in expectation, we would assume 8 of them belong to c. However, this does not hold in modern DL-based models (20), and thus outputs cannot be directly used to define classification thresholds based on desired levels of confidence. ML researchers typically mitigate this limitation by employing a plot of sensitivity versus the complement of specificity (ROC) to select a classification probability threshold providing a reasonable trade-off between these measures for the given task (21). However, medical applications often operate in error regimes in which such an analysis misrepresents relevant clinical performance (13). Worse, a recent meta-analysis in *The Lancet* found that many AI-based medical-imaging papers set thresholds arbitrarily to 0.5 without justification (5). Moreover, even when selecting the optimal threshold in term of ROC, no guarantees are provided regarding the statistical properties of the model's predictions, such as confidence intervals or P values. Although it is generally not possible, in practice, to obtain provable statistical guarantees on performance (as it requires knowledge of the true data distribution), adherence to requirements on metrics such as sensitivity or specificity is nevertheless a critical metric considered in clinical evaluation during prospective studies. Thus, estimating predictions' level of uncertainty and quantifying their statistical robustness are key to cultivating trust among clinicians (1, 19) and may even be more important than improving accuracy (14). Therefore, we define this need as the ability of the system to provide 1) an uncertainty estimation for each prediction and 2) uphold out-of-sample statistical significance requirements specific to the medical task at hand: for example, a model for screening tasks that reports a prediction's SD per person (1) while maintaining a predefined false-positive rate (2).

Heterogeneousness: Handling Data Containing Unknown Medical Conditions or Heterogeneous Mixtures of Conditions

A classification model is typically trained on examples coming from several known classes (e.g., a specific set of medical conditions), in which each example is of a specific class. Given a new example, it assigns a probability to each class, and the class with the highest probability is chosen. When presented with an example from a novel class or an example containing more than one known or unknown class, the model will still output probabilities for the known ones, and the single highest-scoring class will typically be chosen without any indication that, in fact, it "doesn't know" how to classify such an example. This is highly problematic in medical applications, in which it is virtually impossible to train on labeled data with every possible condition in existence or to only encounter real-world data, which is homogeneous in the classes contained within. Here, we refer to cases in which a patient exhibits a condition that the system was simply not trained to recognize, even though other distribution factors may be similar (e.g., the patient comes from the expected population and the data were recorded in the expected way and using similar devices). In a survey of intensive care unit (ICU) and emergency care doctors, practicing clinicians asserted that a model's awareness of situations in which it might be inaccurate or irrelevant is a crucial property in order for it to be useful (14). Consequently, we state this need as the ability of the model to 1) clearly convey to the user that it doesn't know any correct prediction for a given input and 2) correctly detect known classes when the input contains a heterogeneous mix of known and unknown classes.

Relevance: Generalizing while Also Explicitly Identifying Input Samples for which the Model Is Irrelevant

A crucial question that arises regarding any medical AI system is under which circumstances can the system's predictions be trusted and whether those circumstances are explicitly conveyed (1, 12, 14, 22). An ML model is trained by iteratively improving its fitness, in terms of some target function, on the available training data. Meanwhile, the real aim is to obtain beneficial results on all possible data, coming from an unknown underlying distribution. The difference, in expectation, between a model's performance on its training data and on unseen data is known as the generalization gap (23, 24), and the smaller the gap, the better its generalization. Generally, the distribution of the training set may be different from the underlying data distribution (e.g., due to sampling bias), and this difference can be a factor contributing to poor generalization. For example, generalization might suffer when training- and test-time data come from different subject populations, when they are recorded using different medical equipment, when the characteristics of the population change over time, and so on (13). Another generalization issue arises due to training on data from a limited set of patients, which may not be representative of the relevant population. Note that this is in contrast to the heterogeneousness need, which referred to cases with similar data distributions, but in which the model was not trained to recognize some of the classes that exist in the data. Here, the concern is that the model will work, for example, on unseen data recorded with the same equipment but fail to generalize to data recorded with different equipment. In such cases of inadequate generalization, the most crucial concern is, again, that there is no way for a clinician to know whether her model's prediction is valid for the given input (1). We argue that an extra precaution should be taken to address these cases even assuming a system that already addresses the uncertainty need. The danger is that even if true confidence intervals are obtained and validated on the test set, in the wild, samples might yet arrive from different distributions even compared to the available test set ("unknown unknows"). This is especially relevant considering that distributions shift over time, while models are generally not trained continuously (12). Such shifts will render previously obtained uncertainty estimations or confidence intervals unreliable, and more so, over time. Therefore, we define this need as the ability of the model to 1) generalize well across patients and databases while also being able to 2) declare that both its predictions and their uncertainty estimations are irrelevant for a given input and should be ignored due to the input coming from a significantly different data distribution (i.e., beyond its generalization abilities).

In the field of cardiology, ECGs are commonly utilized for diagnostic purposes, therefore making them perfectly placed to leverage DL algorithms' capabilities for the analysis of continuous signals. Applications of ECG analysis include screening and detection of cardiac arrhythmias, which are abnormal heart rates or rhythms. These are high-impact and well-studied applications due to the major global health burden, increased mortality risk, and health costs (25–28) created by cardiac arrhythmias.

We selected two highly clinically relevant (29–31) ECG analysis tasks and systematically addressed the unmet needs with multiple DL-based solutions in order to develop an ECG-analysis system that is clinically useful while delivering results competitive with the current state-of-the-art methods. Specifically, we 1) formulated the issues of clinical relevance in terms of actionable problems to be solved using modern learning techniques; 2) provided explainability of the AI system's decision in terms of the arrythmia-related ECG morphology via a customly-designed input importance weighting mechanism, Spectro-Temporal Attention (STA); 3) showed that our system can uphold statistical-significance requirements and convey confidence levels by incorporating an uncertainty-estimation mechanism directly into it; 4) provided the system with a natural way of outputting any number of known class labels including zero, thus handling inputs coming from unknown classes or even a heterogeneous mix of classes; and 5) supplied an indicator for irrelevance of the system's output in cases of distribution shift induced by utilizing cross-datasets generalization as an intrinsic training objective. We validate our system in a retrospective clinical-screening setting, enforcing an evaluation regime that closely mimics the challenges associated with medical data, such as low train-test samples ratios, data from various populations or recorded with different equipment, data containing more than one medical condition, and other challenges.

Results

Clinical Applications and Approach Overview. In order to assess the benefit of addressing each unmet need and demonstrate its importance for integrating AI into the practice of cardiology, we focus on two ECG-analysis tasks. Firstly, classification of cardiac arrhythmias in heterogeneous-rhythm segments. We trained our model to detect any combination of 10 different prevalent rhythm types from either one- or two-lead ECG segments containing multiple heart beats, which may include a heterogeneous mix of any known or unknown rhythm types and noise. Specifically, the model is tasked with providing binary classification for each rhythm type in each input ECG segment. The second task is detection of whether an underlying pathology is present in a patient, from ECG segments containing regular morphology (i.e., segments in which it is not visible due to the intermittent nature of arrhythmias), thus detecting patients suffering from some known pathologies, which were classified as normal sinus rhythm (NSR) segments by a human cardiologist. To this end, when creating our training data, we make a class distinction between NSR segments from healthy subjects, which we denote simply as NSR, and segments labeled as NSR by a cardiologist but coming from patients with some underlying cardiopathology, which we denote as latent-pathology NSR (LP-NSR). The purpose of this distinction is to train the model to discriminate between these cases, which were indistinguishable to the human cardiologist annotator. The choice of analyzing one or two leads was based on our insistence to work with publicly available data only. Technically, the model can analyze any number of leads, including, but not limited to, the usual 12.

Note that, with the aim of large-scale ECG-based screening in mind, we do not attempt to classify which underlying arrhythmia is present in LP-NSR segments, but to indicate whether such a segment originated in a patient suffering from any arrhythmia. These tasks were chosen because they model what might take place in a clinical environment while a subject is connected to ECG recording equipment for a short duration (minutes). Since cardiac arrhythmias can be highly intermittent, during this time, the ECG may or may not show any known arrhythmias, show any combination of known ones, or show a completely normal rhythm (NSR). Additionally, any number of noise artifacts may be present due to movement or other factors. Thus, the ability of the model to handle heterogeneous and noisy segments is imperative, and the ability to classify a seemingly normal NSR segment as belonging to a subject with an underlying arrhythmia could significantly improve early screening applications (32).

We simulated a clinical setting in both our training and evaluation schemes. Mainly, we cannot assume any prior knowledge on a sample's content; samples will be varied, both in terms of population and recording equipment, and, finally, the inference population is larger than the one used for training. To this end, we incorporated the following guidelines into our methodology: 1) We use a training set substantially smaller than the test set; 2) We do not discard ECG segments for any reason including but not limited to the presence of multiple classes, significant levels of noise, or other characteristics; 3) We include varied data from several devices and population groups in our test set, some of which are deliberately not represented in the training set; 4) The training and test sets are completely disjoint patientwise. See *SI Appendix*, Tables S1 and S2 for the full specifications of the training and test data. Our model's inputs are segments extracted from ECG recordings along with a set of rhythm labels containing the types of rhythms within that segment (see *Materials and Methods*).

We train on half the patients from three openly available datasets from PhysioNet (33): the Normal Sinus Rhythm (33), Long-Term Atrial Fibrillation (34), and MIT-BIH Arrhythmia databases. We define our test set using the remaining patients from these and two additional datasets, the MIT-BIH Atrial Fibrillation (35) and the Telemetric and Holter ECG Warehouse (THEW) (36) databases, which are not represented in the training data. Furthermore, we define an extended test set, which additionally also contains the test set of Hannun et al. (10) and the Computers in Cardiology (CinC) 2017 challenge dataset (37). In total, the extended test set includes 6,584 patients from seven datasets. The train and test set contain data with two ECG leads, while the extended test set contains single-lead data. These datasets all contain ECG data of widely varying duration, with rhythm annotations from expert cardiologists. Moreover, they are highly diverse in terms of subject demographics, recording devices used, rhythm types, and in terms of the imbalances between the different rhythm types. These data contain substantial amounts of missing or noisy segments, which correspond to locations where data were too noisy for a cardiologist to annotate, and which we deliberately did not remove (*SI Appendix*, Table S3).

Fig. 1 presents a bird's-eye view of our proposed framework. Our system is based on a temporal convolution neural-network architecture, which is effective at learning long-range dependencies in time-series data (38). With cardiological tasks in mind, we make some important adaptations and additions to this basic architecture, aimed at addressing each unmet need. Firstly, to gain clinically meaningful explanations, we add a custom input importance weighting mechanism, inspired by Vaswani et al. (39), called STA. STA is designed to exploit both spectral and temporal information, allowing us to pinpoint periodic morphological structures in the input ECG segments–influencing model's decisions the most. Secondly, we employ a multiclassifier architecture in which a separate classifier, subsequently denoted as a "head," is trained per class on shared features extracted by the neural network, in contrast to the standard approach, in which a single multiclass classifier is applied to the learned features. This aids in addressing two of the needs: 1) It allows the model to naturally handle unknown classes by representing such cases as negative classification by all heads, and, furthermore, it allows the model to simultaneously classify a single input into more than one of the possible classes; 2) It enables defining different statistical-performance requirements per class through

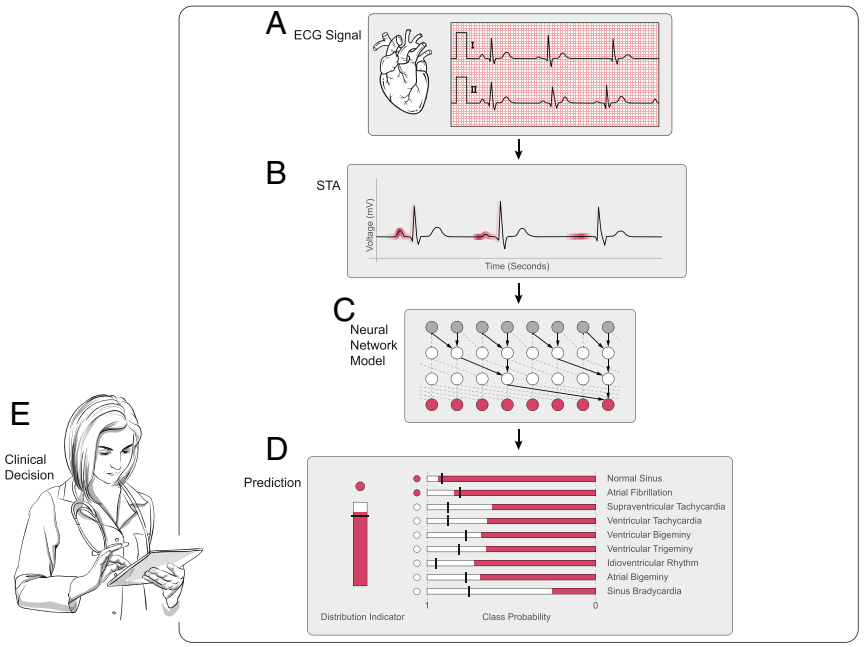

Fig. 1. Schematic of our framework. (*A*) Multiple ECG leads, possibly containing a mix of known and unknown rhythm types, are provided as inputs for the model. (*B*) Our custom STA layer is applied, producing input importance maps highlighting the regions of input contributing the most to the prediction. (*C*) A deep neural network analyzes the attended inputs. (*D*) Separate binary classifiers provide the probability of each rhythm type; different classification thresholds are used for different rhythms based on statistical significance requirements together with a distribution indicator function that expresses whether the model's predictions are relevant for the given input. (*E*) Finally, a clinician chooses whether to trust the model based on the indicator value and uses the rhythm predictions combined with the STA-highlighted regions to make an AI-supported clinical decision. Note that the signals portrayed are only schematic and are not physiological recordings.

control of different per-class classification thresholds. Thirdly, for improving cross-patient generalization, we add a patient classifier to incorporate a surrogate task, which prevents the model from learning patient-specific features (see ablation study in *SI Appendix*, Table S5). Moreover, we introduce an input-relevance indicator by incorporating a dataset classifier, flagging out-of-distribution samples on which the model should not be trusted regardless of its confidence. Finally, we employ a unique classification-threshold selection scheme based on a patientwise disjoint validation set.

Explainability Based on ECG Morphology. Based on our STA layer (see *Materials and Methods*), we derive elementwise attention masks indicating the relative importance of each part of the ECG-input signal to the model's output. Using the STA mechanism, we provide a morphology-based measure of explainability for 10 different rhythm types in both temporal and spectral domains and across any required temporal length. As can be seen in Fig. 2, our system is able to produce justification for the model decisions, expressed as patterns in the signal morphology, similar to how cardiologists analyze ECG. This type of domain-specific explainability is a desirable attribute of medical AI systems (14, 40). Fig. 2*A* displays an ECG sample with an NSR from a patient with no underlying cardiopathology. The model attends each lead in a somewhat different manner, focusing largely on the P and T waves in the upper lead and on the QRS complexes and T waves in the lower lead (blue arrows). Note that a normal morphology of these structures is an indicator of NSR (41).

Fig. 2*B* also displays an ECG sample with an NSR; however, this time, it is recorded from a patient with a confirmed underlying cardiological pathology. In this case, it is apparent that the lower lead contains noise (red arrow) and no relevant information. Importantly, even though the model was not trained to classify noisy samples, the lower lead's attention score shows that the model mostly ignores this lead. This demonstrates that the model with STA is able to detect and ignore ECG segments without viable information content. On the other hand, in contrast to Fig. 2*A*, the upper lead is now endowed with high attention at all major ECG structures (i.e., P wave, T wave, and QRS complexes [blue arrows]). We thus conjecture that the model is able to adapt its information-extraction method in accordance to the specific input. Fig. 2*C* displays an ECG sample with an atrial fibrillation (AF) rhythm. Again, the lower lead is of lower quality, and the model has learned to assign it lower attention. In the upper lead, it can be clearly seen that high attention is consistently assigned to the expected location of P waves (black arrows), which are missing due to the AF. Such behavior is in exact accordance with the cardiological literature and is consistent with how a cardiologist would have explained an AF classification decision (41).

Fig. 2*D* displays an ECG sample with a ventricular tachycardia (VT) rhythm. VT is defined as a heart rhythm higher than 120 beats per minute and is characterized by wide QRS complexes (41). Both indicators can be determined from the duration and location of the QRS complexes. In this case, we can see that the highest attention in both leads is assigned to the QRS complex peaks, as well as to their width, by attending their starting and ending locations (black arrows). Overall, Fig. 2 illustrates how our model is able to consistently give morphological evidence supporting its decisions in a way that is in accordance with the current clinical literature, an ability which was deemed vital for medical AI by clinicians (14).

In addition to these representative examples, we further validated the efficacy of the STA algorithm in a quantitative way on the entire test set. *SI Appendix*, Fig. S1 visualizes the median normalized attention weights produced by STA on five different submorphologies per rhythm class (see methodology in *Materials and Methods*). Moreover, we compared our method to a general-purpose importance-weighting approach, gradient-weighted class activation

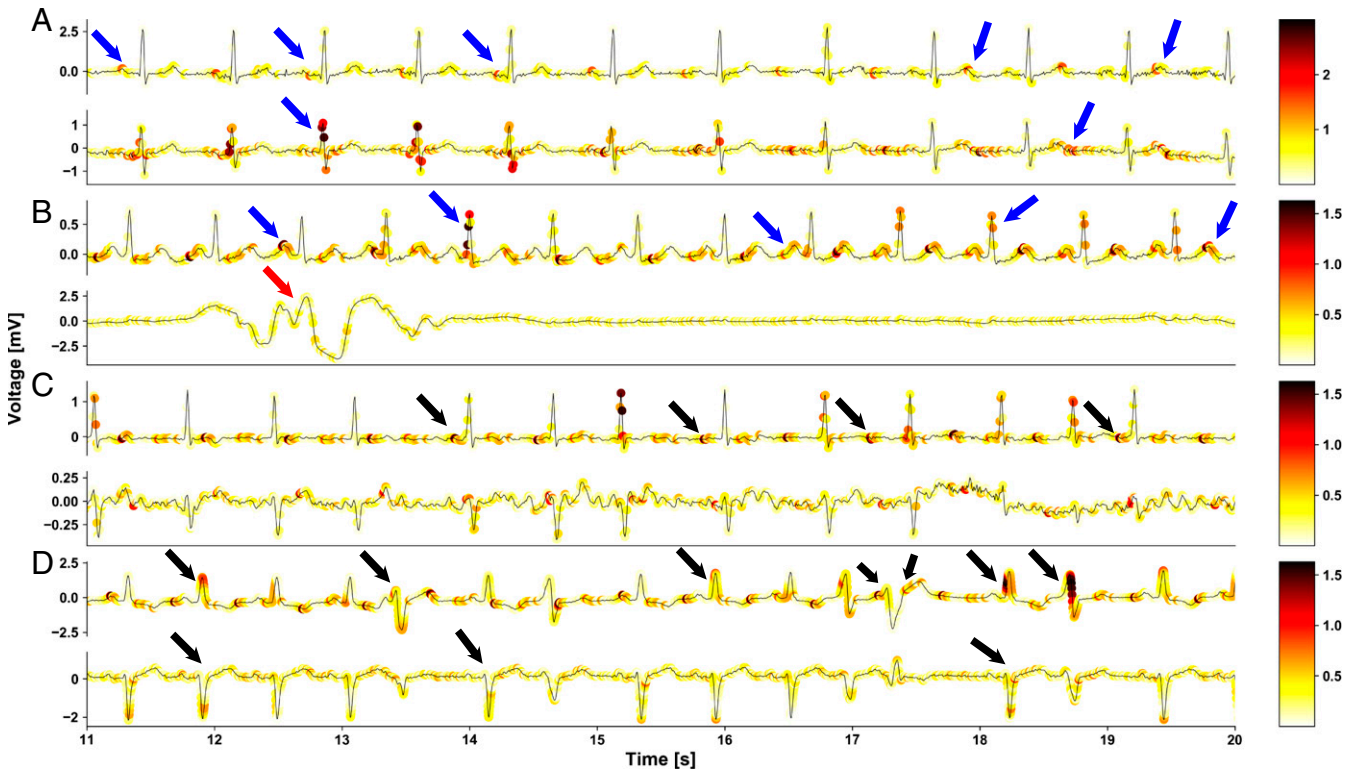

Fig. 2. Temporal attention maps generated by the model for four test set samples, each containing two ECG leads and belonging to a different patient. For clarity, only 9 s are displayed from each sample. (*A*) NSR from a healthy subject with no underlying cardiac pathology. (*B*) NSR from a patient with an existing underlying cardiac pathology (LP-NSR). (*C*) AF rhythm. (*D*) VT rhythm. Blue arrows denote morphologically normal features in normal sinus segments, red arrows denote noise, and black arrows denote abnormal features in ECG segments related to an arrhythmia. These examples showcase the ability of the STA mechanism to detect and highlight periodic components in the input due to the way it is calculated from its spectral representation. Therefore, note that the arrows shown were chosen to present a few exemplary features of interest; they do not represent all relevant high-attention morphological features.

mapping (Grad-CAM) (42), which was recently employed for explainability in ECG (43). The comparison was performed on the test set while normalizing the total attention scores of both methods to the same scale (see *Materials and Methods*). The results (*SI Appendix*, Fig. S2 and Table S4) show that STA gives substantially more attention to clinically salient features, namely specific submorphological segments. STA scores correspond better to a clinically related submorphology for the given rhythm. For instance, in AF and atrial bigeminy, STA gives higher scores to the P wave and PR interval compared to Grad-CAM. For supraventricular tachycardia and VT, which are associated with wide QRS complexes, and for ventricular trigeminy, associated with three consecutive QRS complexed, STA gives a higher attention score to the QRS complex compared to Grad-CAM.

Uncertainty Estimation and Statistical Performance. In order to obtain uncertainty estimates, we applied Monte-Carlo (MC) dropout, proposed by Gal et al. (44). Dropout stochastically disables some percentage of activations of a neural network and is usually applied while training as a form of regularization (45). Here, we use it also during inference and apply the model to the same input multiple times with different dropped activations. This effectively creates a pseudoensemble of different models and provides multiple predictions for each input sample. Gal et al. proved a theoretical link between MC dropout and Gaussian processes (44), showing that the outputs of this pseudoensemble are normally distributed. Therefore, we calculate the SD, per sample and output class, which provides an uncertainty estimate for each prediction. By coupling this method with our multihead classification architecture, we gain separate uncertainty measures for the detection of each rhythm class.

Another crucial aspect relating to uncertainty of predictions is upholding statistical performance requirements: for example, a maximal allowed false-positive rate. Statistical performance depends on the selection of the probability thresholds for positive predictions, which are traditionally chosen by employing ROC analysis for each class. ROC curves describe the relationship between the true-positive rate (TPR), which, in statistical hypothesis-testing terminology, is the statistical power or the complement of the type-II error rate (β) and the false-positive rate (FPR), also known as the type-I error rate (α) or significance level, for different threshold values. Here, our aim was to uphold robust requirements for the type-I error rate in unseen data. To that end, we opted to compute a classification threshold based on a target type-I error rate on a set of validation samples and then compute type-I error and the statistical power at the calculated threshold on a test set of disjoint patients. For the threshold selection, we created the validation set by randomly sampling 20% from the original test set with stratification. We emphasize that no learning steps were performed with the model using the validation set and that only the remaining 80% of the test set were used for the statistical performance evaluation. Table 1 summarizes the statistical performance over both the validation and test sets. We set the type-I error requirement on the validation set to be $\alpha \leq 0.01$. Notably, for all classes, the type-I errors on the test set were either equal to or lower than the target. Furthermore, the system exhibits surprisingly similar statistical power per rhythm between the validation and test sets. These results are likely due to the combination of approaches. The use of separate binary heads together with a pseudoensemble (via MC dropout) produces stable distributions of output probabilities, while performing the threshold selection on a disjoint validation set facilitates generalization of the statistical

performance to unseen data. For further details, see *Statistical Analysis*.

Handling Unknown Rhythm Types and Multirhythm Segments. An immediate result of using separate binary-classification heads for each rhythm type is the ability to systematically deal with unknown classes. By decoupling the prediction task into multiple binary heads, we give the model an intrinsic way to designate an unknown class by outputting a negative prediction in all heads. Moreover, separate classification heads naturally endow our model with the capacity to detect a set of different rhythm types in a single ECG segment.

As shown in Fig. 3, this allows us to train our model with multiple true labels for each segment and prevents the need for a heuristic approach to determine a single-class label in cases in which multiple rhythms are present (e.g., ref. 46), or worse, to discard all nonhomogeneous segments. Fig. 3 displays four exemplary cases. The first is a sample with a single rhythm type, which the model was trained on. This is the main case considered by the majority of prior works. The second is a sample containing two rhythm types, both of which are known to the model. As can be seen, a clinician can be made aware of the presence of both rhythm types in the ECG, since the probability estimations of both rhythms crossed the given threshold. The ability to simultaneously recognize two different classes in the same ECG segment is, by itself, a critical requirement for such systems that is not always addressed, since usually a single input segment is classified to only one of several possible classes (see *Discussion*). The third is a sample containing two rhythm types, one known to the model and the other an unknown rhythm. Although the model obviously cannot recognize an unknown rhythm, it is nevertheless successful in recognizing the known one despite the presence of the other. The fourth is a sample containing only an unknown rhythm type. A traditional classification model would not recognize it is facing such a scenario and would mistakenly output a prediction of one of the known classes. In contrast, our framework naturally handles this situation and indeed outputs probability estimations that are all below the given threshold, causing the predictions for all rhythms to be negative. Hence, the model conveys two crucial pieces of information to the clinician: 1) that the sample in question does not belong to a healthy individual, because there are no positive predictions for either of the NSRs; and 2) the model is unfamiliar with whatever rhythm types are present.

In this way, the system avoids producing misleading results on rhythm classes on which the model was not trained and provides transparency about its limitations. Note that the selected thresholds account for the uncertainty of the model regarding predictions for each rhythm type as well as the statistical significance requirements as determined by the clinician. They are then used jointly with the SD of each specific prediction (obtained from our pseudoensemble). The clinician can thus determine not only whether the prediction has crossed a threshold and is positive for the corresponding rhythm type but also the system's confidence in classifying that specific sample as positive or negative.

Table 1. Statistical performance on the validation and test sets

	Validation			Test		
Rhythm class	Samples	α	$1-\beta$	Samples	α	$1-\beta$
NSR	44,740	0.01	0.81	171,029	0.0006	0.80
LP-NSR	94,092	0.01	0.53	360,432	0	0.53
Atrial fibrillation	142,717	0.01	0.57	545,323	0.01	0.56
Supraventricular tachycardia	4,182	0.01	0.86	16,063	0.009	0.85
Ventricular tachycardia	1,501	0.01	0.97	5,806	0.009	0.96
Ventricular bigeminy	5,110	0.01	0.98	18,942	0.01	0.98
Ventricular trigeminy	1,923	0.01	0.89	7,067	0.01	0.90
Idioventricular rhythm	10	0.01	0.11	56	0.0001	0.01
Atrial bigeminy	5,079	0.01	0.80	19,881	0.009	0.80
Sinus bradycardia	12,186	0.01	0.91	46,636	0.01	0.91

Threshold selection was based on a significance requirement $\alpha \leq 0.01$ on the validation set. α denotes type-I error rate, and $1-\beta$ denotes statistical power. The results show that the significance requirement was met on the test set, even though threshold selection was based on the validation set, with only a minor sacrifice of statistical power in most cases.

Generalizing across Patients and Detecting Data-Distribution Shift. The distribution of medical data observed by an AI model depends on the patients being recorded, the medical equipment used, data storage formats, and possibly other factors. Medical AI systems cannot feasibly be trained with patient data from every conceivable source and must therefore be able to generalize. Our framework is designed for cross-patient generalization by employing regularization through an auxiliary loss function (see *Materials and Methods*) via a domain-adversarial approach (47, 48). Interpatient generalization is improved when the model is trained to extract features that allow rhythm classification but not patient identification. Our results show that learning of patient-specific features is reduced, and thus generalization to new patients is improved, due to this approach (*SI Appendix*, Table S5).

In order to further address the relevance need, we added a dataset-classification layer in order to obtain an indicator function for out-of-distribution samples. In this case, we allow that the model may exploit dataset-specific features so that the indicator enables a clinician to ignore the model if the input is substantially different from samples in the training distribution (e.g., due to population features changing over time). The dataset classifier is optimized separately so as not to affect the feature extraction (see *Materials and Methods*). In order to assess cross-dataset generalization capabilities in a realistic way, our test set deliberately includes data from databases on which we do not train and which were recorded with different types of ECG equipment (*SI Appendix*, Table S1). Although the known and unknown dataset distributions are not completely separated, likely due to the limited capacity of this layer, there is a considerable difference between the two (*SI Appendix*, Fig. S3). The system can utilize this difference for classifying samples as either known or unknown (i.e., as suitable or unsuitable for analysis by the system). This allows clinicians to set a reliability threshold (i.e., an α below which they are not willing to trust the system). For example, for $\alpha = 0.05$, we found that the indicator function includes 40% of samples from known datasets (on which the model was trained and evaluated on disjoint patient sets) and rejects 95% of samples from unknown datasets (on which the model was only evaluated). Both generalization tasks are pursued simultaneously during training.

We performed a comparative analysis to evaluate the importance of employing the indicator function at inference time. The comparison is based on the same methodology of a simulated clinical trial presented in the next section. The results show that when disregarding the indicator function, we observe both a noticeable decrease in overall performance (Table 2 and *SI Appendix*, Table S6) and, more importantly, a failure to uphold the specified statistical-significance requirement (Table 3 and *SI Appendix*, Table S7).

Performance in a Clinical Setting. As elaborated above, the ability of a model to demonstrably uphold statistical performance requirements determined in advance and according to medical considerations is crucial for clinical adoption of AI systems. Therefore, care must be taken when selecting classification thresholds in order to meet such requirements. We opted to evaluate our model under two threshold selection schemes: 1) the ROC optimization scheme (denoted as ROC): Classification thresholds are selected by computing the ROC curve (TPR versus FPR) and choosing a threshold corresponding to the point closest to (0, 1). This scheme is widely used in ML and medical applications. 2) The FPR-constraints

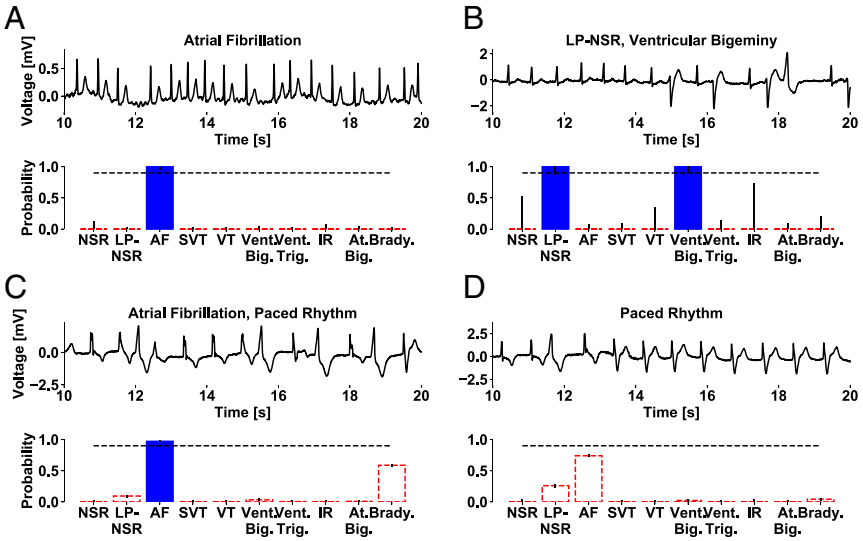

Fig. 3. Multiclass prediction. An input sample to the model is a 30- or 60-s ECG segment, which may simultaneously contain zero or more known types of ECG rhythms as well as unknown types. All the known rhythm types present in the same sample comprise the set of its ground-truth labels. The model outputs a likelihood estimation for each known rhythm type (blue or red bars) along with an uncertainty estimation, which is equivalent to 1 SD in the prediction probability (vertical black lines). A positive prediction for a specific rhythm (filled blue bar) is produced if its likelihood value crosses a predetermined threshold, and a negative prediction is produced otherwise (empty red bar). For visualization only, the threshold was set to the same value of 0.9 for all rhythm types (horizontal black line), though it should generally be chosen per rhythm type according to statistical requirements. Four samples are shown containing one or more known or unknown rhythm types: (*A*) sample containing only one known rhythm (AF); (*B*) sample containing two known rhythms (LP-NSR and ventricular bigeminy); (*C*) sample containing one known (AF) and one unknown (Paced) rhythm, hence there is no output for the unknown type; and (*D*) sample with a single unknown rhythm type (Paced). supraventricular tachycardia, SVT; ventricular bigeminy, Vent. Big.; ventricular trigeminy, Vent. Trig.; idioventricular rhythm, IR; atrial bigeminy, At. Big.; sinus bradycardia, Brady.

scheme (denoted as FPR): Classification thresholds were selected to uphold constraints on type-I errors (FPRs), which correspond to significance levels. We required a significance level of $\alpha = 10^{-4}$ for all rhythm classification heads and $\alpha = 0.05$ for the indicator function. In a clinical setting, these numbers would be based on medical considerations, possibly different per rhythm type. Importantly, for both schemes, the thresholds were calculated using the validation set alone. The motivation for evaluating the system under these two schemes is to study their effect on its statistical performance. While the first scheme (ROC) is most widely employed in the literature, the second scheme (FPR) is arguably more relevant for healthcare providers.

Following threshold calculation, we evaluated all test-set samples under the chosen thresholds from each scheme. This step is akin to a prospective clinical study, which would be performed on the fully trained ML system with predetermined thresholds prior to its deployment. Results are reported in Table 2 for the known test set, and in *SI Appendix*, Table S8 for the extended test set. The requirement of $\alpha \leq 0.05$ for the indicator function results in a significant number of samples being flagged as unsuitable for analysis, resulting in a lower number of analyzed samples per rhythm under the FPR scheme. However, this showcases the ability of our approach to recognize samples that are unsuitable for analysis due to possible distribution shift between the training and test sets. This is in contrast to traditional ML systems, which have no way of detecting such a shift. Based on the consistency of the statistical results reported in Table 1 on both the validation and test sets, we conclude that the results in Table 2 likely reflect the system's would-be real-world performance.

As a concrete example of the importance of threshold selection schemes, we consider a screening task, in which extremely small false-positive rates are necessary due to the desire to apply the screening system to a population as large as possible. We performed threshold selection on the same validation set with a requirement of $\alpha \leq 10 - 4$. The results, showing adherence to the statistical requirements on the test set, are presented in Table 3 and on the extended test set in *SI Appendix*, Table S9. Also considering Table 2, the results show how, despite seemingly superior performance, the model with the ROC scheme generally fails to uphold the prespecified statistical constraints, while the model with the FPR scheme successfully upholds them for all but one rhythm type.

Identification of Background Arrhythmias from NSR Segments. Our second chosen task was to differentiate between two types of ECG segments labeled as morphologically normal by a cardiologist: segments from healthy subjects (NSR) and segments from subjects suffering from a background pathology that does not manifest itself in the specific segment (LP-NSR). Due to the intermittent nature of arrhythmic events, they do not appear in all segments recorded in cardiac patients, and thus such segments are (correctly) labeled as NSR by human annotators. Moreover, because LP-NSR segments are, by definition, segments labeled as normal by cardiologists, this is an example of a task for which the AI system is expected to perform a feat infeasible for a human doctor. We emphasize that the point is to detect whether some pathology may exist in a seemingly normal segment, not which pathology it is. Note also that other ECG-classification works generally do not make any distinction among NSR-labeled samples. For this task, samples were filtered based on the indicator function's score using a naive threshold of 0.5.

We were able to clearly show that 1) almost no LP-NSR samples were classified as NSR (specificity >0.999) and 2) LP-NSR samples are classified overwhelmingly correctly (accuracy is 0.96). Although NSR samples were sometimes classified as LP-NSR (accuracy is 0.70), the results show that the model has substantial predictive power in this task, which, by definition, human cardiologists could not perform. Moreover, the results demonstrate the immense potential of this task for automated population screening, since only three samples from sick subjects (LP-NSR) out of 360,478 were

Table 2. General performance on the test set using threshold selection based on ROC optimization (ROC) or on a significance-level requirement (FPR) of $\alpha \leq 10^{-4}$

	Analyzed samples		Accuracy		AUC		Specificity		Sensitivity		Precision		F1 score	
Rhythm class	ROC	FPR	ROC	FPR	ROC	FPR	ROC	FPR	ROC	FPR	ROC	FPR	ROC	FPR
NSR	170,779	20,823	0.96	0.99	0.89	0.93	0.99	1.0	0.78	0.99	0.99	0.87	0.87	0.93
LP-NSR	360,309	87,880	0.76	0.65	0.78	0.73	0.69	0.59	0.89	0.88	0.59	0.35	0.71	0.50
AF	545,430	321,451	0.81	0.61	0.84	0.73	0.96	0.99	0.66	0.48	0.95	0.99	0.78	0.64
SVT	16,024	578	0.99	0.99	0.91	0.95	0.99	0.99	0.66	0.80	0.85	0.86	0.74	0.83
VT	5,787	146	0.99	1.0	0.90	0.95	0.99	1.0	0.80	0.88	0.85	1.0	0.83	0.93
Vent. big.	19,044	5,283	0.99	0.99	0.95	0.99	0.99	0.99	0.90	0.99	0.80	0.90	0.90	0.90
Vent. trig.	7,141	456	0.99	0.99	0.90	0.93	0.99	1.0	0.76	0.83	0.83	0.96	0.79	0.89
IR	54	12	1.0	1.0	0.52	0.50	1.0	1.0	0.02	0.0	1.0	0.0	0.03	0.0
At. big.	19,736	877	0.99	0.99	0.89	0.97	0.99	0.99	0.66	0.82	0.87	0.74	0.75	0.78
Brady.	46,832	8,873	0.99	0.99	0.92	0.97	0.99	0.99	0.86	0.95	0.88	0.94	0.87	0.94

Note that the exact values presented are not the major point of interest. Instead, we emphasize the substantial difference between the ROC and FPR schemes, demonstrating the need to apply a proper threshold selection scheme in order for the system to adhere to required statistical constraints. Although naively optimizing on the ROC appears to yield superior performance in some cases based on common metrics, the FPR-based threshold selected using the validation set obtains comparable results while also adhering to statistical requirements on the test set (Table 3). AUC, area under the ROC curve; SVT, supraventricular tachycardia; Vent. big, ventricular bigeminy; Vent. trig., ventricular trigeminy; IR, idioventricular rhythm; At. big., atrial bigeminy; Brady., sinus bradycardia.

misclassified as healthy (NSR). The complete results are available in *SI Appendix*, Table S10.

Discussion

AI has the potential to deliver radical breakthroughs in medical applications. However, the question of how to bridge the gaps between the promising results presented in AI research papers and the real-world clinical setting must be addressed in order to reap any potential benefits. These gaps, which we have formulated as the unmet needs of clinicians from AI systems, pertain to multiple different issues, which must all be addressed together in order to achieve a viable system that can be deployed in clinical practice. We defined each of the needs based on concerns raised by clinicians and in a way actionable by AI researchers. We then demonstrated this actionability by addressing each unmet need for high-impact clinical tasks in AI-based ECG analysis.

Explainability. We showed that interpretability, at least in the sense defined by clinicians, can indeed be engineered into a DL model for ECG-analysis tasks. Our STA mechanism was specifically designed to detect periodic components and then highlight them in the temporal domain. We have shown that it provides a visualization of the relative importance of each part of an ECG segment for the final model decision. Thus, interpretability is provided in terms of the morphological features present in each specific ECG segment, which corresponds to the way cardiologists are trained to analyze ECG and explain their judgments. For example, a cardiologist may point to abnormal or missing P waves in a patient's ECG in support of a diagnosis of AF. Likewise, our STA mechanism highlights the most relevant ECG regions for its decision (Fig. 2).

In other cases, the regions highlighted by the STA can shed light on subtle morphological details a human cardiologist is likely to miss. This was demonstrated by our second task, in which underlying cardiopathology was successfully detected in segments that had normal morphology according to a cardiologist (LP-NSR). We do not claim that highlighting morphological details should provide a cardiologist with novel clinical insight regarding a pathology, especially when it is intermittent and does not manifest in the observed ECG segment. However, to the extent that useful morphological features exist in a specific segment, the goal of explainability is achieved because the evidence for the model's results is communicated in a way a cardiologist can reason about. Moreover, our system was able to produce explanations for its decision, which often conformed to established medical knowledge about morphological features of specific arrhythmias. In a recent work published in *The Lancet* (32), researchers from Mayo Clinic demonstrated an ability to detect the presence of AF in patients by analyzing NSR segments of ECG, thus establishing the clinical viability of detecting pathology from morphologically normal segments. Although our system does not attempt to classify which underlying arrhythmia is present from NSR segments (unlike in ref. 32), it does indicate whether an NSR segment is abnormal with respect to multiple other rhythm types. Consequently, it is arguably more useful for large-scale automated ECG-based screening.

Table 3. False-positive rates on the test set using threshold selection based on ROC optimization (ROC) or on a significance-level requirement (FPR) of $\alpha \leq 10^{-4}$

Scheme	NSR	LP-NSR	AF	SVT	VT	Vent. big.	Vent. trig	IR	At. big.	Brady.
ROC	10^{-4}	0.31	0.03	$2 \cdot 10^{-3}$	$7 \cdot 10^{-4}$	$2 \cdot 10^{-3}$	10^{-3}	0	10^{-3}	$5 \cdot 10^{-3}$
FPR	10^{-5}	0.41	10^{-5}	$5{\cdot}10^{-5}$	10^{-5}	$4 \cdot 10^{-5}$	$6{\cdot}10^{-5}$	0	10^{-5}	10^{-5}

Thresholds were selected based solely on a validation-set (disjoint patient-wise from the train set). The model with ROC scheme fails to uphold the required FPR for all but one rhythm (IR), while the model with the FPR scheme successfully upholds the requirement for all rhythms but LP-NSR, which the model occasionally confuses with NSR (recall these are morphologically identical for cardiologists). A failure to uphold the statistical requirement is highlighted in red. SVT, supraventricular tachycardia; Vent. big, ventricular bigeminy; Vent. trig., ventricular trigeminy; IR, idioventricular rhythm; At. big., atrial bigeminy; Brady., sinus bradycardia.

Previous works, such as Mousavi et al. (49) and Yao et al. (50), have applied attention in the temporal domain in order to gain a measure of explainability in morphological terms. However, these works have not been able to show clear correlation between known morphological features and the inputs on which their models relied the most in their decisions. Other recent works have shown more success in providing explainability for DL-based ECG models with emphasis on clinical relevance. Raghunath et al. (43) used a general-purpose method, which produces visual explanation in convolutional neural networks (CNN) models, Grad-CAM (42), and applied it to ECG analysis. Meira et al. (51) generated textual explanations for an existing model based on the duration of submorphological features known to clinicians, such as P waves, QRS complexes, etc. We combined ideas from both of these works and compared our STA algorithm to Grad-CAM for each submorphology and rhythm type. Our comparison is based on the entire test set, not on representative examples, and shows the median of normalized attention scores generated by each method (*SI Appendix*, Fig. S2 and Table S4). The results indicate a substantial difference in distribution of the attention scores to submorphologies between these methods. The STA median and maximal normalized scores are consistently higher for submorphologies, which are considered to be clinically salient and therefore provide better explainability. In contrast, the Grad-CAM approach places more of its attention on parts of the ECG outside of these submorphologies. Furthermore, the STA scheme provides per-lead attention scores (*SI Appendix*, Fig. S1), while Grad-CAM does not. Considering that cardiologists and some DL-based systems (52) use 12-lead ECGs to inspect different morphological features, the importance of per-lead explanation provides another advantage of STA. These results highlight the explainability benefits of designing an attention mechanism explicitly tailored to ECG. Importantly, this work reports a quantitative comparison between two different explainability methods for ECG on the grounds of best adherence to clinically relevant morphological features, a comparison which is usually not provided.

Uncertainty Estimation. We have incorporated uncertainty estimation directly into our model inference stage. In a clinical setting, this allows us to simulate a model ensemble for each sample to be classified and thus calculate the SD of the predictions based on the ensemble outputs as an uncertainty measure. We press further with this approach and show how we can allow a clinician to select the classification threshold separately for each arrhythmia type in a statistically rigorous way based on the required significance level (Table 1). Furthermore, this work shows a DL-based medical AI system that is able to uphold predetermined statistical constraints on unseen data, a critical quality, which is largely missing from most other similar works. Thus, the proposed threshold-selection scheme can be employed by clinicians using DL-based systems under rigorous statistical requirements specified by medical needs. Many prior works specifically address the issue of estimating predictions' uncertainty. In addition to the approach of Gal et al. (44), which we adopted, a prominent and more recent example is Romano et al. (53), in which the authors were able to obtain theoretically guaranteed confidence intervals without access to the true underlying distribution. Although remarkable, these guarantees stem from an assumption regarding the interchangeability of samples from the train and test sets. In most medical tasks, however, samples from patients in the train set are differently distributed than samples from patients in the test set and different still from the ones later encountered in the wild; thus, the assumption does not hold in practice.

Heterogeneousness. We have shown that our multihead architecture inherently solves this issue by simply providing the model with a natural way to not apply any classification to an ECG segment. Indeed, our results show that the model was able to achieve this [e.g., for segments that only contained unknown rhythm types (Fig. 3)]. Other approaches for detecting unknown classes exist in the literature, such as adding a "background" class, training a separate classifier to detect unknown classes (54), or defining loss functions that allow detection of an unknown class based on thresholding the maximal softmax output (55). These approaches are largely orthogonal and can be incorporated with our approach. We opted for the multihead design in this regard because it also allowed us to naturally frame our classification tasks as a set-level operation (i.e., we detect the set of all rhythm types in an ECG segment instead of classifying each input as only one of the possible rhythms). In contrast, in many other works dealing with detection of arrhythmias such as AF (46, 56, 57) and other pathological conditions such as congestive heart failure (58) or myocardial infarction (59), the task is framed as a binary classification: either the analyzed segment is classified with the specified condition or not. We argue that such a binary framing is inherently unsuitable for clinical applications because it does not allow the model any way to proclaim its inadequacy in case the data do not belong to either the positive or negative class it was trained with or to account for samples displaying a mixture of conditions. In a survey of ICU and emergency care doctors, practicing clinicians asserted that a model's awareness of situations in which it might be inaccurate or irrelevant is a crucial property in order for it to be useful (14). We believe that by combining our model's ability to provide an uncertainty estimation for each prediction, its ability to not classify the segment as any known class, and its ability to declare its output as invalid via the indicator function, we have thoroughly addressed the concerns raised by the clinicians from this survey.

Relevance. Our solution addresses the issue of cross-patient generalization as well as the problem of input distribution shift due to factors such as different demographics or medical equipment. We address these issues separately and with a different approach. By combining domain-adversarial training, which considers the patients as domains with an indicator function for the data distribution, we substantially mitigate issues of generalization to new patients (*SI Appendix*, Table S5) and datasets. In our simulated clinical setting, our model was able to handle data from patients it was not trained on (the common case), and the distribution indicator function could notify the clinician in case the model's outputs are not usable in this case, thereby significantly improving the clinical viability of the system (*SI Appendix*, Tables S6 and S7). In this regard, we wish to note the inherent trade-off between training a decoupled dataset classifier (our indicator function) versus adversarially training it (as with the patients' classifier). While adversarial training will optimally result in dataset-agnostic learned features and thus will have likely improved generalization, it will also hinder any capability to judge whether samples originated from a known distribution seen at training. However, instead of opting for better generalization in the dataset case, our method provides a quantifiable metric by which to judge whether a sample is suitable for processing in the first place. This is akin to the concept known as selective prediction, in which a separate classifier is used in order to decide whether the main classifier's output should be considered for an input. However, in contrast to many works in this area, we train the indicator function together with the model. A notable recent example of this approach is that by Geifman et al. (54), which also incorporates a custom loss function to train the indicator function. We opted for a simplified approach, in which we train the indicator as a classifier for different domains and control the coverage using the classification threshold (*SI Appendix*, Fig. S3). Based on opinion papers, such as those by Rajkomar et al. and Tonekaboni et al. (1, 14, 19), we deem the benefits of such a function outweigh those of improved generalization in a clinical setting. Other prominent journal publications, such as those by Hannun et al. (10) or Attia et al. (60), claim the clinical viability of their DL-based ECG-analysis solutions. However, these authors do not explicitly attempt

to address generalization in the design of their models and instead opt to demonstrate it on a different dataset or a held-out test set from the same dataset. Moreover, according to a survey by Luz et al. (61), many other ECG-classification works from recent years fail to apply even the most basic measure for cross-patient generalizability: interpatient training (i.e., using a disjoint set of patients) for train, validation, and test time.

Regarding our chosen method for addressing each unmet need, we wish to emphasize that, with the exception of the STA mechanism, our approach is not specific to ECG analysis and, in fact, not even to medical tasks. In our view, explainability in medical tasks should be domain specific by design; thus, we demonstrated this by supporting the model's predictions in terms of ECG morphological features, which are salient to cardiologists (40). However, all other components of our system are applicable to any DL-based approach. Specifically, they could be employed for clinically relevant AI systems in non-ECG– or non-cardiology–related tasks.

In conclusion, we have demonstrated a medical-AI system that addresses the unmet needs defined above in the context of ECG-based classification tasks while also performing favorably on two high-impact tasks validated on a challenging and realistic mix of datasets. Therefore, we believe our framework has a number of substantial benefits with respect to its clinical applicability in comparison with previous work in the domain of ECG analysis. However, notwithstanding these successes, we stress that our goal was to clearly define and overcome the notable gaps between clinical needs and medical-AI research, not to advocate our specific AI system as ready for deployment. In this light, our suggested set of unmet-needs represent the demands posed by clinicians in a way that allows consideration by designers of medical-AI systems, and our proposed solutions should be viewed as an established reference for addressing them. Thus, we believe this work represents a significant step forward in understanding and overcoming the limitations that currently impede wider adoption of AI-based systems in both cardiology and medicine in general.

Materials and Methods

Data Preprocessing. To achieve a realistic training setting, only the following four minimal preprocessing steps were applied: 1) We removed 5 min from the beginning and end of each recording in order to discard segments of electrode placements or removals.* This step was performed because many long ECG recordings contain no signal at the start or end of the recording due to electrode placement or removal. We emphasize that, in a clinical setting, this step will not be required, and, in principle, a signal of any length can be processed. 2) We detrended the remaining signal using a second-order polynomial. This step was necessary due to common low-frequency oscillations in the recordings and would not be necessary for the evaluation of a single short recording. 3) Extreme values of each ECG signal were clipped, with extreme values defined as being over six SDs away from the mean. 4) We scaled the signal to the dynamic range of [−1,1] by dividing by its maximal absolute value. The clipping was done in order to normalize the extremity of noise peaks without completely removing them, and the value of six SDs was empirically chosen as sufficiently extreme. We stress that, after these preprocessing steps, no additional ECG segments were discarded for any reason from either our training or validation datasets. Note that all of the above-mentioned four preprocessing steps are performed only for the benefit of the optimization process. At no point would the physician be exposed to the normalized signals, as the attention maps are displayed on the raw, unnormalized signal produced by the ECG recording device.

Creating Labeled Samples. Each ECG record was divided into overlapping windows containing $B \in \{30,60,90\}$ beats, with an overlap of 26, 56, and 86 beats, respectively. We then segmented and resampled each beat on a constant sampling grid of $K - 1$ samples centered around the beat's R peak. This was performed for each of $L \in \{1,2\}$ ECG leads in the record. We set $K = 82$ in all cases. In addition, we calculated a $(K - 1)$-point estimate of the power spectral density (PSD) for each beat using Welch's method (61) and the R-peaks intervals (RRI) vector. The B ECG beats from each window and each ECG lead were stacked side-by-side columnwise into a [K, B] matrix, where the first row of the matrix contains the RRI vector. The motivation for including the RR interval with each beat was to keep information regarding the heart-rate variability, a meaningful feature of cardiac function (62, 63) that was lost during the resampling process, since each beat is now represented by an identical number of points, with the R peaks being always set to the middle point of its respective column. Similarly, PSD matrices of shape [K, B] were constructed, with a zero vector in the first row. The beat and PSD matrices from all leads were concatenated along a new first dimension, thus producing a three-dimensional tensor of shape [2L, K, B] for each window, where L is the number of ECG leads.

Each such tensor was considered as a single input sample for our model, with the corresponding assigned label being the set of rhythm type annotations for all of the contained beats. Note that resampling around the R peak and stacking beats in this manner effectively aligns their representations on a stable fiducial point.

Model Architecture. The model is composed of three parts (Fig. 1): 1) STA, a custom method that aligns between input channels from the time and frequency domains and provides insight into the model's predictions; 2) a deep temporal convolution neural network (TCN), which extracts nonlinear features from an input tensor; and 3) C + 2 classification heads, with C being the number of different rhythm types the model is trained to recognize. Each one of the C heads is a two-layer fully connected (FC) neural network with the first layer shared across all heads. The two additional heads are used for patients and dataset classification and contain a single FC neural network each.

STA. The STA mechanism was in part inspired by the self-attention mechanism (39) used across multiple tasks that require modeling temporal dependencies within a sequence, such as machine translation (39) and image generation (62). STA is based on the idea of self-attention but is specialized to incorporate both time- and frequency-domain representations of the input signal. The motivation for STA is twofold; first, it allows our convolutional model to take advantage of spectral information in the input. Since spectral data do not share common support axes with temporal data, it would be meaningless to simply concatenate these representations and apply to them a local convolutional filter. Thus, STA allows influence to flow between multiple spectral and temporal samples, effectively by learning a transformation aligning these complementary signal representations. Second, by learning to create this alignment per ECG window, we obtain a visualization method that accentuates the relationship between the domains. This increases the model's interpretability by highlighting the most relevant temporal and spectral features for each prediction. Thus, our STA module was designed to model the interdependence between the time- and frequency-domain representations of a signal and, specifically, to learn a weighting, or "attention," of time-domain features (i.e., signal morphology) based on the frequency-domain representation of the signal (i.e., power spectral density), and vice versa. Full details and mathematical formulations are available in the *SI Appendix*.

TCN Architecture. A diagram of the architecture is presented in *SI Appendix*, Fig. S4. Our TCN was designed broadly following the approach of Bai et al. (38). This architecture takes advantage of dilated convolutions in order to increase receptive-field size. However, instead of one-dimensional causal convolutions, we employ two-dimensional noncausal ones, since causality is not relevant for this type of task: We process a full ECG segment as one sample, and we use its full temporal extent for producing an output for the entire segment, not for each timestep within. Overall, our TCN is composed of five repeating blocks, each containing nine metalayers: two of type A, two of type B, and one of type P. These metalayers contain one or more convolutional layers and are defined as follows:

- Type A (1): Dilated convolution layer, $k = 4 \times 2$, $d = 2$ (2); Convolution layer, $k = 5 \times 3$, $s = 1$ (3); A residual connection between the input and the output of the unit, containing a convolution layer, $k = 1 \times 1$, $s = 1$.
- Type B (1): Dilated convolution layer, $k = 4 \times 2$, $d = 2$ (2); Convolution layer, $k = 5 \times 3$, $s = 1$ (3); A direct residual connection between the input and the output of the unit.
- Type P: Strided convolution layer, $k = 5 \times 3$, $s = 2$. Note that this unit type serves as a pooling operation in our network. For the 30-beat-per-window model, we reduced the stride of the last three P units to 1 to prevent the spatial extent on the signal being reduced to zero generalizable ones. An ablation study for γ is presented in the *SI Appendix*.

*This was not performed on data from the CinC2017 challenge or the Hannun et al. test set (10, 37), in which the duration of each sample is between 30 and 60 s.

The parameters k, s, and d denote kernel size, strides, and the dilation level, respectively. The order of the metalayers within each block is A, B, A, and B repeated twice, followed by P. Following each convolutional layer in any unit, we employ batch normalization (63), a nonlinear activation, and dropout regularization (45) with a drop probability of d, with d = 0.4 in all experiments. The nonlinear activations are all of type LeakyReLU (64) with $\alpha = 0.2$. The final block output serves as the input to all classification heads.

Classification Heads. We define C + 2 classification heads implemented as FC layers. C heads are binary and are used for detecting rhythm classes. The two additional nonbinary heads predict the patient ID and the database from which the sample was obtained for the purpose of interpatient generalization as described in the sequel. All C rhythm classification heads have 2,048 hidden units and also share an initial FC layer with 2,048 hidden units.

Auxiliary Confusion-Loss Functions. In order to encourage interpatient generalization, we adapted the approaches introduced in (47, 48) and derived a method that both improves interpatient generalization and provides an indicator function for out-of-distribution samples. Both goals are obtained via a similar mechanism.

We add two single FC layers whose inputs are the outputs of the TCN module. These layers are trained to identify the specific patient ID and the database from which the sample originated. To induce an adversarial confusion loss on our model, we employ a gradient-reversal layer (47) for both classification heads, which flips the sign of the gradient coming from the classification heads during backpropagation and applies a factor of γ to its magnitude. For the patient classification head, we set $\gamma = 0.1$. In essence, this penalizes the model's feature extractor parts for the success of the patient classifier. Consequently, the confusion loss regularizes the STA and TCN modules to not extract patient-specific features but instead to be more robust. Pursuing a distribution indicator function, we trained the dataset classification head with $\gamma = 0$. This does not regularize the feature extractors but instead decouples the database classifier head's training from that of the larger model. We then use it to estimate the probability of a new sample being sampled from a data distribution that was previously seen by our model.

Statistical Analysis. In order to specifically limit the type-I statistical error of the system during inference, we have randomly sampled a validation test (as described in *Results*). We then computed the empirical cumulative distribution function (CDF) of scores from each rhythm classification head for all positive validation samples of the relevant rhythm, C_{pos}, as well as the empirical CDF of all negative validation samples of the relevant rhythm, C_{neg}, from which we then computed the complementary $C = 1 - C_{neg}$. Note that for each threshold τ, it holds that $\alpha \leq C(\tau)$, where α denotes the type-I error of the model on the validation set. We can also compute the power of the model on the validation set, where $power = 1 - \beta$, $\beta = C_{pos}(\tau)$, and β denotes the type-II error of the model on the validation set. The scores used for computing the empirical CDFs were the average scores obtained from the dropout-based pseudoensemble.

For a given minimal required α, we choose a score threshold τ s.t. $\alpha = C(\tau)$, which is then used as a classification threshold for the specific rhythm type for which the CDF was computed. For indicator-function threshold selection, we employ the exact same method, with the single exception of replacing the positive classification score of each rhythm head with the maximal probability score given by the indicator function across all known databases.

STA Visualization and Comparison to Grad-CAM. In order to visualize the STA results on the entire test set, and also in order to quantitatively compare between our STA method and an existing method, we used Grad-CAM, which was previously used in the literature for importance weighting in the context of ECG, with the following methodology. We performed the comparison on the test set data (with which the results displayed in Table 2 were also calculated). Each sample in this dataset is a two-lead ECG segment containing 60 beats, with different rhythm annotations known per beat. We calculated the attention scores generated by both methods and normalized them so that the total attention of each method on the entire ECG segment (all 60 beats together) is equal to 100. We then used the ecgpuwave package available from PhysioNet (33) to automatically annotate each ECG segment and mark all onsets and ends of the following submorphologies and subsegments: P wave, PR interval, QRS, ST interval, and T wave. We only used the submorphological parts that ecgpuwave was able to successfully detect. For each submorphology segment in each ECG sample, we also obtained the rhythm annotation from the ECG beat it belongs to. We summed the attention score of all points contained in each submorphology of each beat to obtain the total attention score for that submorphology in that specific beat. From this, we calculated aggregate statistics (such as the median value) of the attention scores per submorphology and rhythm type. Additional notes include the following: 1) points that are not contained in the submorphology segments are not used in these visualizations; therefore, the total visualized attention score is not 100. 2) Visualized scores are for a single submorphology in a single beat, whereas the normalization is across 60 beats; therefore, the values do not represent a percentage of attention for each submorphology but rather a relative score to compare between different submorphologies.

Data Availability. All study data are included in the article and/or *SI Appendix*.

ACKNOWLEDGMENTS. This work was supported by the Israel Ministry of Science (A.S. and Y.Y.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This research was partially supported by the Technion Hiroshi Fujiwara Cyber Security Research Center and the Israel Cyber Directorate (Y.Y. and A.M.B.)

1. A. Rajkomar, J. Dean, I. Kohane, Machine learning in medicine. *N. Engl. J. Med.* **380**, 1347–1358 (2019).
2. J. He *et al*., The practical implementation of artificial intelligence technologies in medicine. *Nat. Med.* **25**, 30–36 (2019).
3. B. Norgeot, B. S. Glicksberg, A. J. Butte, A call for deep-learning healthcare. *Nat. Med.* **25**, 14–15 (2019).
4. A. Esteva *et al*., A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
5. X. Liu *et al*., A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *Lancet Digit. Health* **1**, e271–e297 (2019).
6. S. M. McKinney *et al*., International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
7. A. Esteva *et al*., Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
8. J. De Fauw *et al*., Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
9. A. Rajkomar *et al*., Scalable and accurate deep learning with electronic health records. *NPJ Digit. Med.* **1**, 18 (2018).
10. A. Y. Hannun *et al*., Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* **25**, 65–69 (2019).
11. E. J. Topol, High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
12. T. Panch, H. Mattie, L. A. Celi, The "inconvenient truth" about AI in healthcare. *NPJ Digit. Med.* **2**, 77 (2019).
13. J. Wiens *et al*., Do no harm: A roadmap for responsible machine learning for health care. *Nat. Med.* **25**, 1337–1340 (2019).
14. S. Tonekaboni, S. Joshi, M. D. McCradden, A. Goldenberg, What clinicians want: Contextualizing explainable machine learning for clinical end use arXiv [Preprint] (2019). https://arxiv.org/abs/1905.05134v2 (Accessed 7 August 2019).
15. B. Goodman, S. Flaxman, European Union regulations on algorithmic decision making and a "right to explanation." *AI Mag.* **38**, 50–57 (2017).
16. Z. C. Lipton, The mythos of model interpretability. *ACM Queue* **16**, 1–28 (2018).
17. D. Chen *et al*., Deep learning and alternative learning strategies for retrospective real-world clinical data. *npj. Digit. Med.* **2**, 1–5 (2019).
18. D. Castelvecchi, Can we open the black box of AI? *Nature* **538**, 20–23 (2016).
19. T. Ching *et al*., Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15**, 20170387 (2018).
20. C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, "On calibration of modern neural networks" in 34th International Conference on Machine Learning ICML 2017, D. Precup, Y. W. Teh, Eds. (*Proceedings of Machine Learning Research*, 2017), pp. 2130–2143.
21. T. Fawcett, An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**, 861–874 (2006).
22. M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, R. Ranganath, A review of challenges and opportunities in machine learning for healthcare arXiv [Preprint] (2018). https://arxiv.org/abs/1806.00388 (Accessed 5 December 2019).
23. Y. Jiang, D. Krishnan, H. Mobahi, S. Bengio, "Predicting the generalization gap in deep networks with margin distributions" in International Conference on Learning Representations (International Conference on Learning Representations, 2019), pp. 1–19.
24. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (MIT Press, 2016).
25. M. Haissaguerre, E. Vigmond, B. Stuyvers, M. Hocini, O. Bernus, Ventricular arrhythmias and the His-Purkinje system. *Nat. Rev. Cardiol.* **13**, 155–166 (2016).
26. L. H. Ling, P. M. Kistler, J. M. Kalman, R. J. Schilling, R. J. Hunter, Comorbidity of atrial fibrillation and heart failure. *Nat. Rev. Cardiol.* **13**, 131–147 (2016).
27. A. S. Adabag, R. V. Luepker, V. L. Roger, B. J. Gersh, Sudden cardiac death: Epidemiology and risk factors. *Nat. Rev. Cardiol.* **7**, 216–225 (2010).
28. D. Mozaffarian *et al*.; Writing Group Members; American Heart Association Statistics Committee; Stroke Statistics Subcommittee, Executive summary: Heart disease and stroke statistics–2016 update: A report from the American Heart Association. *Circulation* **133**, 447–454 (2016).

29. G. H. Mairesse *et al*.; ESC Scientific Document Group, Screening for atrial fibrillation: A European Heart Rhythm Association (EHRA) consensus document endorsed by the Heart Rhythm Society (HRS), Asia Pacific Heart Rhythm Society (APHRS), and Sociedad Latinoamericana de Estimulación Cardíaca y Electrofisiología (SOLAECE). *Europace* **19**, 1589–1623 (2017).
30. B. Freedman *et al*.; AF-Screen Collaborators, Screening for atrial fibrillation. *Circulation* **135**, 1851–1867 (2017).
31. D. H. Birnie, W. H. Sauer, M. A. Judson, Consensus statement on the diagnosis and management of arrhythmias associated with cardiac sarcoidosis. *Heart* **102**, 411–414 (2016).
32. Z. I. Attia *et al*., An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: A retrospective analysis of outcome prediction. *Lancet* **394**, 861–867 (2019).
33. A. L. Goldberger *et al*., PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **101**, E215–E220 (2000).
34. S. Petrutiu, A. V. Sahakian, S. Swiryn, Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans. *Europace* **9**, 466–470 (2007).
35. G. B. Moody, R. G. Mark, "A new method for detecting atrial fibrillation using RR intervals" in *Proceedings of 10th Computers in Cardiology (CinC)* (IEEE, 1983), pp. 227–230.
36. J. P. Couderc, The Telemetric and Holter ECG Warehouse initiative (THEW): A data repository for the design, implementation and validation of ECG-related technologies. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2010**, 6252–6255 (2010).
37. G. D. Clifford *et al*., AF classification from a short single lead ECG recording: The PhysioNet/computing in cardiology challenge 2017. *Comput. Cardiol.* **44**, 1–4 (2017).
38. S. Bai, J. Z. Kolter, V. Koltun, (2018) An empirical evaluation of generic convolutional and recurrent networks for sequence modeling arXiv [Preprint] (2018). https://arxiv.org/abs/1803.01271v2 (Accessed 19 April 2018).
39. A. Vaswani *et al*., *"Attention is all you need" in Advances in Neural Information Processing Systems,* I. Guyon, Ed. *et al*. (Nips, 2017), vol. 2017, pp. 5999–6009.
40. D. Kasper *et al*., *Harrison's Principles of Internal Medicine* (McGraw-Hill, ed. 19, 2015).
41. D. L. Mann, D. P. Zipes, P. Libby, R. O. Bonow, E. Braunwald, *Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine*, E. Braunwald, R. O. Bonow, Eds. (Elsevier/Saunders, Philadelphia, ed. 10, 2015).
42. R. R. Selvaraju *et al*., "Grad-CAM: Visual explanations from deep networks via gradient-based localization" in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (IEEE, 2017), pp. 618–626.
43. S. Raghunath *et al*., Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nat. Med.* **26**, 886–891 (2020).
44. Y. Gal, Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning" in *International Conference on Machine Learning*, M. F. Balcan, K. Q. Weinberger, Eds. (Proceedings of Machine Learning Research, 2016), pp. 1050–1059.
45. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
46. S. P. Shashikumar, A. J. Shah, G. D. Clifford, S. Nemati, "Detection of paroxysmal atrial fibrillation using attention-based bidirectional recurrent neural networks" in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18* (ACM Press, New York, 2018), pp. 715–723.
47. Y. Ganin *et al*., Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **17**, 1–35 (2016).
48. E. Tzeng, J. Hoffman, K. Saenko, T. Darrell, "Adversarial discriminative domain adaptation" in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. 2017-January (IEEE, 2017), pp. 2962–2971.
49. S. Mousavi, F. Afghah, A. Razi, U. R. Acharya "Ecgnet: Learning where to attend for detection of atrial fibrillation with deep visual attention". in *IEEE EMBS International Conference on Biomedical & Health Informatics* (IEEE, 2019), pp. 1–4.
50. Q. Yao, R. Wang, X. Fan, J. Liu, Y. Li, Multi-class arrhythmia detection from 12-lead varied-length ECG using attention-based time-incremental convolutional neural network. *Inf. Fusion* **53**, 174–182 (2020).
51. W. Meira, A. L. P. Ribeiro, D. M. Oliveira, A. H. Ribeiro, Contextualized interpretable machine learning for medical diagnosis. *Commun. ACM* **63**, 56–58 (2020).
52. V. Gliner *et al*., Automatic classification of healthy and disease conditions from images or digital standard 12-lead electrocardiograms. *Sci. Rep.* **10**, 16331 (2020).
53. Y. Romano, E. Patterson, E. J. Candès, Conformalized quantile regression. arXiv [Preprint] (2019). https://arxiv.org/abs/1905.03222 (Accessed 1 March 2021).
54. Y. Geifman, R. El-Yaniv, "SelectiveNet: A deep neural network with an integrated reject option" in *International Conference on Machine Learning*, K. Chaudhuri, R. Salakhutdinov, Eds. (Proceedings of Machine Learning Research, 2019).
55. A. R. Dhamija, M. Günther, T. E. Boult, Reducing network agnostophobia arXiv [Preprint] (2018). https://arxiv.org/abs/1811.04110 (Accessed 1 April 2020).
56. S. Mousavi, F. Afghah, ECGNET: Learning where to attend for detection of atrial fibrillation with deep visual attention. arXiv [Preprint] (2018). https://arxiv.org/abs/1812.07422 (Accessed 1 April 2020).
57. N. Keidar, Y. Elul, A. Schuster, Y. Yaniv, Visualizing and quantifying irregular heart rate irregularities to identify atrial fibrillation events. *Front. Physiol.* **12**, 637680 (2021).
58. M. Porumb, E. Iadanza, S. Massaro, L. Pecchia, A convolutional neural network approach to detect congestive heart failure. *Biomed. Signal Process. Control* **55**, 101597 (2020).
59. U. R. Acharya *et al*., Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals. *Inf. Sci.* **415–416**, 190–198 (2017).
60. Z. I. Attia *et al*., Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat. Med.* **25**, 70–74 (2019).
61. E. J. d. S. Luz, W. R. Schwartz, G. Cámara-Chávez, D. Menotti, ECG-based heartbeat classification for arrhythmia detection: A survey. *Comput. Methods Programs Biomed.* **127**, 144–164 (2016).
62. H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, Self-attention generative adversarial networks. arXiv [Preprint] (2018). https://arxiv.org/abs/1805.08318 (Accessed 1 April 2020).
63. S. Ioffe, C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift" in *32nd International Conference on Machine Learning, ICML 2015*, F. Bach, D. Blei, Eds. (Proceedings of Machine Learning Research, 2015).
64. K. Xu *et al*., "Show, attend and tell: Neural image caption generation with visual attention" in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, F. Bach, D. Blei, Eds. (Proceedings of Machine Learning Research, 2015), pp. 2048–2057.