Automatic Detection of Non-Biological Artifacts in ECGs Acquired During Cardiac Computed Tomography

Rustem Bekmukhametov[1], Sebastian Pölsterl[1(✉)], Thomas Allmendinger[2], Minh-Duc Doan[2], and Nassir Navab[1,3]

[1] Computer Aided Medical Procedures, Technische Universität München, Munich, Germany
{r.bekmukhametov,sebastian.poelsterl,nassir.navab}@tum.de
[2] Diagnostic Imaging and Computed Tomography, Siemens Healthcare GmbH, Forchheim, Germany
{thomas.allmendinger,minh-duc.doan}@siemens.com
[3] Johns Hopkins University, Baltimore, MD, USA

Abstract. Cardiac computed tomography is a non-invasive technique to image the beating heart. One of the main concerns during the procedure is the total radiation dose imposed on the patient. Prospective electrocardiographic (ECG) gating methods may notably reduce the radiation exposure. However, very few investigations address accompanying problems encountered in practice. Several types of unique non-biological factors, such as the dynamic electrical field induced by rotating components in the scanner, influence the ECG and can result in artifacts that can ultimately cause prospective ECG gating algorithms to fail. In this paper, we present an approach to automatically detect non-biological artifacts within ECG signals, acquired in this context. Our solution adapts discord discovery, robust PCA, and signal processing methods for detecting such disturbances. It achieved an average area under the precision-recall curve (AUPRC) and receiver operating characteristics curve (AUROC) of 0.996 and 0.997 in our cross-validation experiments based on 2,581 ECGs. External validation on a separate hold-out dataset of 150 ECGs, annotated by two domain experts (88 % inter-expert agreement), yielded average AUPRC and AUROC scores of 0.890 and 0.920. Our solution is deployed to automatically detect non-biological anomalies within a continuously updated database, currently holding over 120,000 ECGs.

Keywords: Anomaly detection · Cardiac computed tomography · Electrocardiography · Prospective ECG gating

1 Introduction

Computed tomography (CT) is a non-invasive imaging technique, where a number of X-ray projections, taken from different angles, form a volumetric image of an area inside the body. Here, we focus on images of the heart, i.e., cardiac CT, which is often used to detect coronary artery disease or to evaluate the heart's

B. Berendt et al. (Eds.): ECML PKDD 2016, Part III, LNAI 9853, pp. 193–208, 2016.
DOI: 10.1007/978-3-319-46131-1_24

function and morphology [9]. Due to constant beating of the heart, cardiac CT is particularly challenging: to ensure sharp motion-free images, multiple X-ray projections need to be taken at the same cardiac phase. In addition, the imaging protocol needs to be optimized to reduce the total radiation dose a patient is exposed to, thereby lowering the risk of radiation-induced cancer [9]. Hence, keeping a proper balance between low radiation exposure and image quality is one of the major trade-offs in a cardiac CT [9].

One of the most effective imaging techniques in this field is based on prospective ECG gating [10], the central idea of which is to activate the X-ray source only at the "right" time windows, namely during the cardiac phases of interest. Such gating algorithms reduce radiation by over 70 %, while maintaining high image quality [9]. On the other hand, relying on ECG makes the whole cardiac CT workflow highly dependent on the quality of the ECG signal, which is influenced by various factors specific to a patient, hospital, and physician. If the ECG signal is corrupted by noise or artifacts, prospective ECG gating is prone to fail and the resulting image of the heart will be of poor quality. In some cases, the scan has to be repeated, which offsets the advantage of prospective ECG gating in reducing radiation dose. We describe typical non-biological artifacts that may disrupt the imaging workflow in Sect. 2.

While the field of ECG analysis is well-established, it addresses problems distinct from ours. Common use case are clinical decision support and patient monitoring, both of which use ECG to assess a patient's health status. Consequently, anomalies of biological origin are the primary focus. In contrast, this work aims to accurately identify ECG signals that are corrupted by various non-biological artifacts, disregarding any medical conditions a patient might have. In addition, the characteristics of ECG signals and artifacts encountered in cardiac CT differ from those encountered in clinical diagnosis (see Sects. 2 and 3). To the best of our knowledge, this is the first scientific work that thoroughly investigates methods to automatically identify anomalies occurring in the context of cardiac CT.

We developed a system that can process large pools of data from multiple medical centers across the world and automatically identify CT scanners experiencing anomalous behavior. Our approach has several advantages. First, it dramatically reduces the time and effort of identifying problems compared to a human analyst, which leaves more time to fix a particular problem. Second, our customers benefit by reduced response times to an incident. Third, we expect that our system helps to increase the rate of high quality cardiac CT images, while maintaining a low radiation exposure. Our solution utilizes existing techniques used in ECG analysis and incorporates two feature extraction methods, which are based on robust PCA [3] and a discord discovery algorithm [13]. We retrospectively analyzed 2,581 cardiac CT scans from 60 medical centers from 18 countries. We evaluated our solution by cross-validation and by comparing its predictions to annotations of two domain experts on a hold-out set of 150 scans. The results demonstrate that our system is highly discriminatory and allows processing thousands of ECGs with minimal human interaction.

The paper is structured as follows. In Sect. 2 and 3 we will describe the most prevalent noise patterns encountered in the context of cardiac CT and our dataset. Section 4 describes our system. Next, we present our evaluation results in Sect. 5 and end with concluding remarks in Sect. 6.

2 Noise Patterns in Cardiac CT

In this section, we will describe the most prevalent noise patterns encountered in cardiac CT. But first, let us provide a brief insight into the ECG signal's morphology. The letters P, Q, R, S, and T name the key features of an ECG waveform. A typical heartbeat starts with a so-called P wave, continues with a QRS complex – characterized by a narrow spike called R peak – and ends with a T wave (see Fig. 1). Each feature corresponds to a particular phase in the cardiac cycle.

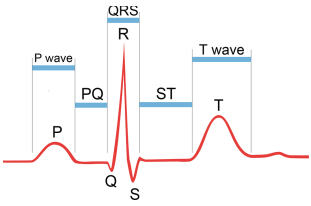


Fig. 1. Typical ECG waveform describing the heartbeart of a healthy patient.

Prospective ECG gating relies on detection of R peaks to predict the time of future R peaks. Since the R peak occurs at a distinct phase during the cardiac cycle, its detection enables imaging the heart in a predefined cardiac phase [9]. However, the presence of noise or non-biological artifacts in the ECG signal may result in false positive R peaks, which, in turn, may cause desynchronization of the whole workflow, resulting in a low quality image and the need for a repeat scan.

Typical non-biological artifacts observed in an ECG during cardiac CT can be classified into the following 6 categories:

- **Powerline noise** is caused by the interference of an ECG signal with an external power supply (Fig. 2a), for instance, if a power cord is placed across the patient or close to an ECG electrode.
- **Baseline wandering** is typically caused by breathing and movement of the patient, and becomes particularly strong when cardiograph's electrodes are unreliably connected to the body (Fig. 2b).
- **Rotational noise** is caused by an electrostatic charge near to or within the scanning area, which results in a rapid change of the electric field formed by the local static charge and rotating high-voltage generators of the CT scanner (Fig. 2c). Note that the noise is eliminated once the scanning process begins, because the X-ray leads to a discharge.
- **X-ray artifacts** are usually due to an X-ray beam hitting a piece of metal. This may happen when an electrode moves in the scanning area or the patient has one or more implants (Fig. 2e,f).
- **Table motion artifacts** are characterized by a noticeable fall of the ECG signal quality while the examination table is moving. Localized baseline and high frequency disturbances are sometimes observed after the table starts moving due to movements of the patient, improper wiring of the electrodes, or other reasons (Fig. 2g).

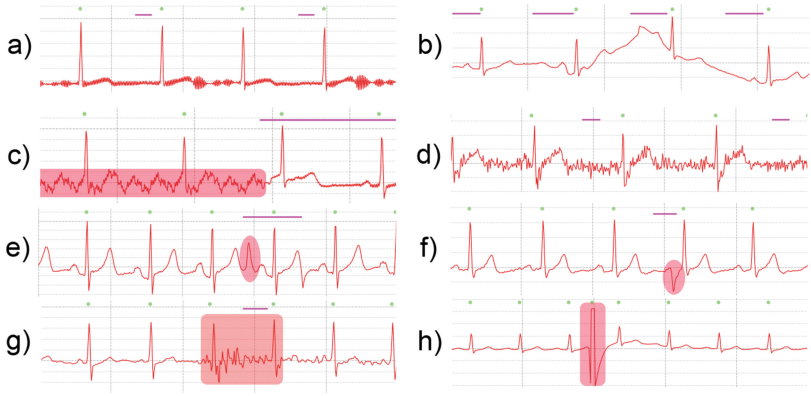


Fig. 2. Noise patterns observed during a cardiac CT scan. (a) powerline interference, (b) baseline wandering, (c) rotational noise, (d) other ubiquitous noise, (e-f) X-ray artifacts, (g) table motion artifact, and (h) localized disturbance. The purple lines indicate time intervals, during which the X-ray scanner was active.

– A wide range of **other noise** types due to a variety of reasons that are either unknown or do not fall into the abovelisted categories (Fig. 2d, h).

Although noise can be minimized by calibrating the CT equipment, not all medical centers may follow the best practices. By proactively identifying potentially unsuccessful scans, we mitigate the aforementioned health concerns and improve customer experience.

3 Dataset

Our dataset comprised 2,581 ECG signals from 60 medical centers from 18 countries annotated by a human expert as either "good" or contaminated with one or more of the noise patterns described above. Each ECG signal was sampled at a frequency of 100 Hz and on average ranged between 30 and 40 seconds in duration. In addition, each trace contained information about the time intervals, where the X-ray source was activated and the positions of QRS complexes, estimated by a proprietary R peak detection algorithm during image acquisition. Analyzing the dataset was challenging due to the following properties:

– ECG signals were highly heterogeneous due to different equipment used, different physicians performing the scan, and different technical and professional standards among countries.
– The ECG signal consisted only of the recording from a single lead, in contrast the conventional 12 lead ECG for clinical diagnosis.
– In cardiac CT, electrodes are placed outside of the patient's chest to not interfere with the X-ray scanner, which often results in atypical ECG waveforms, where only the R peak can be identified reliably.

– Relatively short ECG recordings, with the average length of 23 cardiac cycles and the minimal length of only 5 cycles.

4 Methods

In this section, we present our system to automatically quantify non-biological artifacts and noise in ECG signals. First, we present a high-level overview of our system. Subsequently, we explain our feature extraction technique for describing anomalous ECG signals. Finally, we illustrate how these features were incorporated into an ensemble of classification models.

4.1 High-Level Overview

Our general strategy is to first split the overall problem into multiple subproblems and address each individually, before combining our separate solutions into a unified system, which yields probabilistic scores representing the magnitude of noise in a given ECG trace. We can formulate two subproblems based on characteristics of the noise patterns depicted in Fig. 2:

1. *Global noise patterns* comprise disturbances that, once present, tend to contaminate the whole signal. This category includes baseline wandering, powerline interference, rotational noise caused by electromagnetic interference, and a subset of other noise types (Fig. 2a-d).
2. *Localized noise patterns* comprise non-biological artifacts that affect the ECG signal only within certain, relatively short, time intervals. It includes X-ray artifacts, disturbances related to movement of the examination table, and other miscellaneous localized disturbances (Fig. 2e-h).

For each category, we develop a feature extraction method tailored to that particular subproblem and train an ensemble of classification models on top of the extracted features to distinguish anomalous ECGs from normal ECGs and to quantify the extent of noise in a trace (see Fig. 3). Our approach can be summarized as follows.

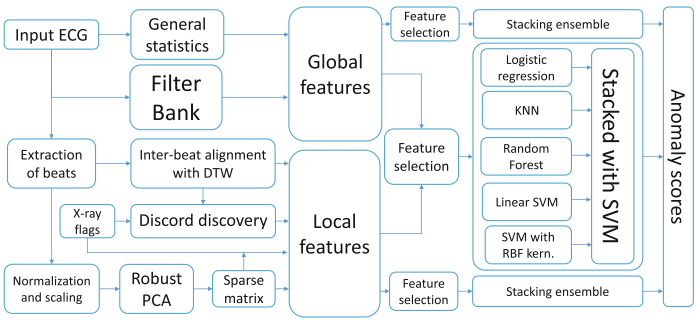


Fig. 3. High-level overview of our system.

1. A filter bank extracts features describing global noise patterns.
2. Each ECG trace is decomposed into a set of non-overlapping intervals constituting a full cardiac cycle – referred to as *beat* – based on the provided locations of QRS complexes.
3. Each beat is analyzed by a modified discord discovery algorithm [13], which identifies the most unusual beat based on the dynamic time warping distance [17] and a test for outliers [16].
4. At the same time, beats are combined into an inter-beat matrix, which is supplied to robust PCA [3] to detect anomalous patterns within this matrix.
5. Next, we compute features describing localized noise patterns based on the output of the previous two steps, i.e., discord discovery and robust PCA.
6. Finally, we use the features describing global and localized noise patterns to train three different ensembles of various classification models, each yielding an anomaly score in the interval $[0; 1]$:
   - The 1st model is trained to exclusively recognize global noise patterns. Its score represents the extent of global noise in a trace.
   - The 2nd model quantifies the extent of localized noise patterns in a trace.
   - The 3rd model, called *unified model*, is trained on the union of features describing global and localized noise patterns. It ought to quantify the overall amount of noise, disregarding the category of noise.

Let us now present individual steps in more detail.

4.2 Global Noise Patterns: Filter Bank Approach

Global noise patterns (Fig. 2a-d), by definition, should be detectable by looking at general properties of the signal. A straightforward approach would consider the signal-to-noise ratio of an ECG signal. Typically, it is estimated as the ratio of the signal's power (P_{signal}) to the power of the noise (P_{noise}):

$$\text{SNR} = 20 \cdot \log_{10} \left(P_{\text{signal}} / P_{\text{noise}} \right).$$

Obviously, we are unable to estimate the SNR in such a straightforward manner, as we do not know the noise component or the reference signal in advance. Instead, we develop a set of filters that separate the noise component from the observed signal. The extracted noise signal can subsequently be used to compute the SNR and to extract other features describing the signal. Next, we compose a set of features that describe the characteristics of global noise patterns.

To filter out powerline interference and baseline wandering (Fig. 2a,b), we utilize that both noise patterns are characterized by certain frequency bands, which would be either absent or much less explicit in unaffected signals. We employ a two-pass median filter [5] to extract noise stemming from baseline wandering, and a notch filter [18] to capture noise due to powerline interference. For clean signals, the extracted noise signal would be negligible and the denoised signal would largely correspond to the original signal.

Separating the remaining noise types is more challenging due to a large overlap between frequencies of the true (biological) signal and the noise. Standard

band-pass filters affect both the noise and the actual signal, thereby distorting the ECG waveform, in particular the QRS complexes. We observed that many artifacts in the rotational noise and "other global noise" category (Fig. 2c,d) resembled white Gaussian noise, which can be filtered efficiently by utilizing the wavelet shrinkage technique [6], which works as follows. First, using the discrete wavelet transform [15], we represent the signal as a weighted sum of basis functions with different time and frequency resolutions. The weights or coefficients of basis functions corresponding to high frequency signals tend to capture the white Gaussian noise, which can be eliminated by applying the soft thresholding operator to the wavelet coefficients and reconstructing the signal via the inverse transform [6]. The result is a denoised version of the input signal with well preserved morphological features of the ECG waveform, in particular R peaks. In addition, we compute the median absolute deviation of wavelet coefficients at the highest resolution level, which quantifies noise based on the wavelet coefficients itself [6]. The features derived from wavelet coefficients and the denoised signal ought to differentiate clean signals from signals affected by rotational noise and various global noise patterns (Fig. 2c,d).

Up to this point, we addressed baseline wandering, powerline interference, rotational noise, and other types of ubiquitous noise independently. We combine these individual approaches into a filter bank: the separated noise and signal components are used to estimate the SNR and the original signal and its frequency domain representation to compute a number of statistics (normalized max. and min. amplitudes, mean, variance, skewness, kurtosis, and entropy). In total, we compute 65 features describing global noise patterns.

4.3 Localized Noise Patterns: Considerations

In contrast to global noise patterns, localized noise patterns (Fig. 2e-h) are characterized by pointwise, temporal changes in the signal, which requires methods operating at a high temporal resolution. Most existing work on ECG analysis is related to clinical diagnosis [11] and human identification [1]. For clinical diagnosis, feature extraction should focus on aspects that characterize a disease and at the same time account for the natural variability of ECG waveforms and heart rhythms across patients. For human identification, features need to differentiate individuals, while mitigating factors that vary across multiple measurements for the same individual, such as heart rate and signal quality. In both applications, the key morphological features of the ECG waveform, such as P wave, T wave, QRS complex, and so forth, often convey sufficient information about diseases and individuals. In our case, we require features that are robust to variations across CT scanners, imaging protocols, and individuals and their diseases. Most importantly, the source of noise patterns considered here is almost always independent from the individual and her heartbeat characteristics. Consequently, standard features of an ECG waveform may not be reliable in our context.

The notion of a localized noise pattern implies that there is a part of the signal, which notably deviates from the rest of the ECG. This suggests to first identify the most anomalous subsequences within the signal and then to assess

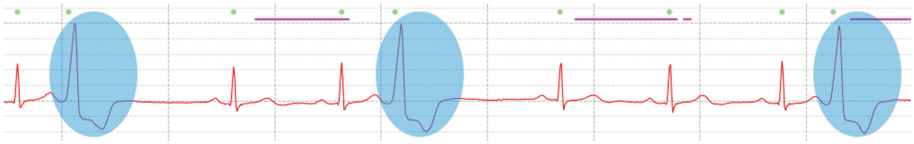


Fig. 4. ECG of a patient with a preliminary ventricular contraction (PVC).

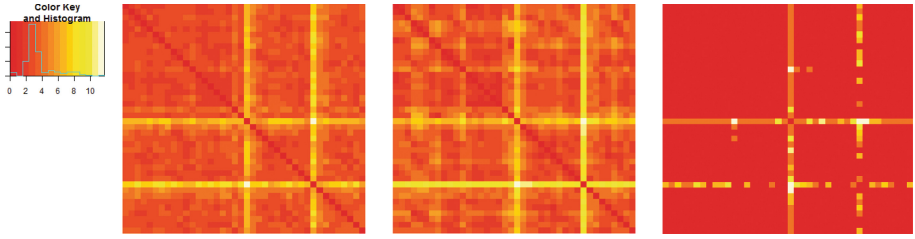
type and degree of these anomalies. One of the key challenges is that such anomalies can be caused by technological as well as biological factors. A biological anomaly in the ECG signal is due to a physiological condition of the heart, such as preliminary ventricular contraction (PVC) depicted in Fig. 4. Therefore, artifacts of biological origin usually occur during specific cardiac phases and repeat themselves over time. In contrast, a technological anomaly, such as a sudden discharge caused by an X-ray, is not associated with a specific cardiac phase, instead, it can occur during any phase. Furthermore, the unique waveform of a technological artifact rarely occurs more than once in the same ECG trace, i.e., it is an outlier. Only in severe cases, when the noise results in a falsely detected R peak, we observe not one, but two beats with an unusual morphology.

Next, we present two feature extraction methods that detect anomalous structures within a signal, while mitigating the natural variability across diseases, patients, and medical centers. The first approach is based on a discord discovery algorithm, which finds the most unusual subsequence within a time series. The second approach utilizes robust PCA for identifying anomalous structures within a signal. These methods operate at the beat level, i.e., the time between two R peaks. We argue that this is the most reasonable level of detail for three reasons: (1) it provides necessary and sufficient information to identify repetitive and anomalous subsequences; (2) as mentioned in Sect. 3, only R peaks are well preserved across all signals; and (3) those traces, where an R peak was misdetected, are usually contaminated with non-biological artifacts and we have ways to recognize them, which we will describe next.

4.4 Localized Noise Patterns: Discord Beat Discovery

Discord beat discovery (DBD) performs a series of comparisons of ECG beats to identify the most anomalous beats. First, the ECG signal is decomposed into multiple beats based on the detected QRS complexes. Next, the beats are normalized to uniform length and compared with each other using a suitable distance measure. The result is an inter-beat dissimilarity matrix (see Fig. 5).

One of the key aspects of this approach is the choice of an appropriate distance measure. The two primary criteria for choosing a distance metric are the ability to handle ECGs with variable waveform morphology and its runtime performance. The latter criteria is crucial, because we require $O(B^2)$ comparisons for each ECG signal, where B is the number of beats in the signal, and we want to analyze thousands of ECG traces in a short amount of time. The Euclidean



(a) Dynamic Time Warping (b) Euclidean distance (c) SAX

Fig. 5. Inter-beat dissimilarity matrices based on different distance measures. The yellow bands correspond to ECG beats contaminated with an X-ray artifact. SAX: symbolic aggregate approximation [14].

distance is fast to compute, but by definition is not tolerant to temporal inconsistencies within time series [17]. The dynamic time warping (DTW) algorithm [17] accounts for such differences and thereby mitigates natural morphological variabilities within the ECG waveform. The symbolic aggregate approximation (SAX) metric performs dimensionality reduction prior to distance comparison and is robust to changes in waveform morphology too [14]. Empirically, we have found that DTW provides the most optimal trade-off between accuracy and computational complexity for our purposes.

Our DBD approach can be considered as a modification of the brute force discord discovery (BFDD) algorithm [13]. The BFDD algorithm often performs well in finding the most unusual part of a time series, but has a runtime complexity of $O(L^2)$, where L denotes the length of the time series. Moreover, it requires us to specify the length of the anomaly, which is rarely known in advance. Instead, we adapt the BFDD algorithm with the following change: instead of comparing all possible subtraces of a fixed length with all others, we split the signal into non-overlapping beats first, and only compare beats with each other. This modification has the following consequences:

1. Focusing on the comparison between beats eliminates the need to specify a fixed window length and better suits the ECG analysis context.
2. Significantly faster runtime of $O(B^2)$ – the number of beats B is about two orders of magnitude smaller than the number of samples L in a signal.
3. It allows for an integration of domain knowledge in the form of predefined patterns (discussed below).

Identifying localized noise patterns can be challenging when the ECG signal contains both technological and biological anomalies, such as PVC beats (Fig. 4). Our DBD approach accounts for the presence of biological anomalies by utilizing that they tend to reappear over time. Therefore, for each beat – regular or biologically abnormal – we can find another beat within the trace whose similar (its distance is small). In contrast, a beat corrupted by a technological artifact possesses a unique waveform – it will have a large distance to all other beats in the trace. The DBD approach is suitable, given the following two conditions:

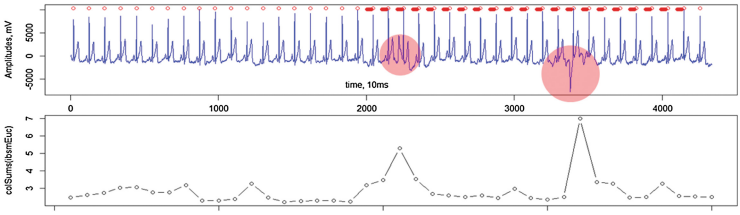


Fig. 6. Example of the discord beat discovery approach. Time intervals, where the X-ray source is active, are marked with red bars. Bottom: distance to the closest beat.

(1) there are at least two biologically abnormal beats, and (2) at least one of them occurs outside of time intervals where the X-ray source was active. These preconditions are met in most scenarios. In a few rare cases, it may happen that there is only one biological anomaly and it appears exactly under an X-ray region, in which case we might assume the presence of a non-biological X-ray artifact. Our solution for such cases is to maintain a set of patterns of typical biological anomalies. Each discord with a statistically significant deviation [16] from its closest beat should be additionally aligned with these patterns. In case a close match is found, the anomaly is likely of biological origin; otherwise, the observed discord is either a technological anomaly or a novelty, i.e., an unexpected form of a biological artifact. In either case, the trace would be of interest for the analyst. Considering the computational overhead of this approach and the rarity of these cases, we decided not to include predefined patterns in our deployed system, but this remains as a potential future improvement.

Figure 6 illustrates the DBD algorithm on an ECG contaminated with several X-ray artifacts. The main disadvantage of the algorithm is that it can only determine the presence of an anomalous beat, but not the exact location and structure of the anomaly. In the next section, we describe a technique that overcomes these limitations.

4.5 Localized Noise: Robust PCA

In this section, we present a technique based on robust PCA [3] that allows for a very precise localization of an anomaly at the sub-beat level, which is particularly useful for capturing X-ray artifacts (Fig. 2e,f). Moreover, this approach allows extracting anomalous structures and reconstructing the true, noise-free signal.

Robust PCA [3] is a modification of classical PCA designed to handle strong outliers. It seeks a decomposition of a matrix X into two components, $X = L+S$, such that L is a low-rank matrix that comprises regular patterns within the data, and S is a sparse matrix, which captures irregular structures. There are no strong assumptions about the irregularities – the only requirement is that they appear unusual with respect to the rest of the data, such that the sparsity condition of the S matrix holds. This enables us capturing a wide range of anomalies. The objective function of robust PCA consists of two terms: the nuclear norm

$\|X\|_* = \sum_i \sigma_i(X)$, with $\sigma_i(X)$ denoting the i-th singular value of X; and the L_1 norm $\|X\|_1 = \sum_{ij} |X_{ij}|$. The resulting optimization problem has the form:

$$\min_{L,S} \quad \|L\|_* + \lambda \|S\|_1, \qquad \text{subject to} \quad L + S = X, \lambda > 0.$$

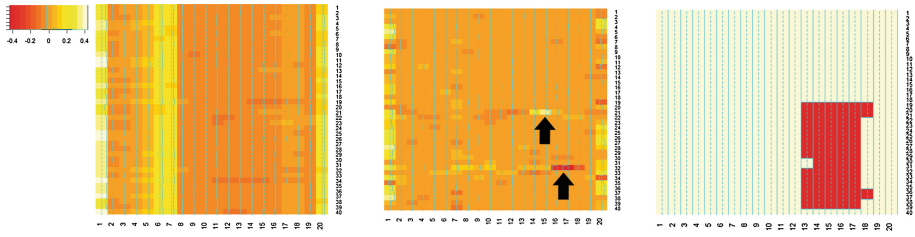
In many applications, the data can be modeled as a sum of such low-rank and sparse components [3]. In the context of cardiac CT, we want L to capture the biological signal and S arbitrary localized non-biological anomalies.

Here, the rows of matrix X correspond to the beats in a single ECG trace: first, we segment the ECG into beats and scale individual beats to uniform length; next, the beats are stacked to form the inter-beat matrix X. Applying the robust PCA procedure to X yields the decomposition into the matrices L and S (see Fig. 7 a and b). An ECG waveform that is severely corrupted by a localized noise pattern, such as an X-ray artifact, can have its true waveform captured by the L matrix and the non-biological anomaly by the S matrix. As a result, the information we are interested in tends to accumulate in the S matrix.

We use S to answer two questions: (1) whether the ECG contains a significant anomaly and (2) whether this anomaly only occurs when the X-ray source is active. First, we apply a test for outliers [16] to the values of S and compute noise quantification measures such as the median absolute deviation. Next, we build a binary X-ray matrix M by splitting the vector of X-ray flags according to R peaks (Fig. 7), and compute the correlation between values in S and the positive flags in M. Finally, we divide S into groups corresponding to different values of the binary X-ray matrix M and perform t-tests to determine whether their mean significantly differs from each other.

We identified three requirements for this approach to yield good results:

1. There are enough beats within the trace to infer repetitive structures (15 beats are usually sufficient).
2. There is either a single beat corrupted by a non-biological artifact or multiple corrupted beats, each with its own unique waveform.
3. Artifacts of biological origin, if present, do repeat over time.



(a) Low-rank matrix L. (b) Sparse matrix S. (c) X-ray matrix M.

Fig. 7. The low rank (a) and the sparse (b) components produced by the robust PCA procedure applied to the ECG signal in Fig. 6, as well as the binary matrix of X-ray flags (c). Two anomalies are captured by the S matrix (marked by arrows). Overlaying (b) and (c) reveals that both anomalies are X-ray artifacts.

Our empirical findings and cross-validation results, which will be presented below, suggest that these three conditions hold for most of the noise patterns. Considering the third condition, in some cases an ECG signal contains few biologically abnormal beats (<15 %) such that the robust PCA mistakes them for outliers, i.e., they are captured by the S matrix. To address this ambiguity, we use our discord beat discovery approach described above, which can handle cases with two to three ectopic beats. By combining features extracted during discord beat discovery and robust PCA, we end up with over 100 features describing localized noise patterns.

4.6 ECG Trace Classification

After developing features describing global and localized noise patterns, respectively, we obtained two sets of features. These sets adequately describe the noise patterns in Fig. 2, but are applicable to their respective domain only and only some of them, such as SNR, directly provide information about the magnitude of noise in a given ECG trace. Therefore, we employ a classification model that utilizes all 181 extracted features to yield a probabilistic score representing the magnitude of noise, global or local. We refer to this model as the *unified model*.

Moreover, we noticed a considerable redundancy in the feature set and that some features contribute little to the overall model. Thus, prior to training, we rank features by importance – using the improvement in out-of-bag error estimated from a random forest (RF) model [2] – and retain all features in the top half. Subsequently, multiple classification models are trained on the selected features to distinguish good from anomalous ECGs. We use RF [2], linear SVM, SVM with RBF kernel [4], k nearest neighbors classification [8], and logistic regression. Although a single model trained on the selected feature set can provide satisfactory results (see Table 1), each model has its own biases determined by its learning principle and its hyper-parameter configuration. Thus, to further raise the reliability of our system, we construct an ensemble of the above mentioned models using the *model stacking* technique [19], where an SVM with RBF kernel is used as meta-model. This increases the complexity of training, but we believe this is acceptable, because the model is rarely re-trained once deployed (the additional costs during prediction are negligible).

Analogous, we train two additional ensembles to recognize global and localized patterns exclusively. Three anomaly scores in the range $[0; 1]$ form the output. The main score is produced by the unified model and represents the final conclusion about the quality of the ECG, because it is equally sensitive to the presence of global and localized noise patterns. The remaining two scores exclusively quantify the amount of global and localized noise, respectively.

5 Evaluation

We evaluated our solution using cross-validation and a hold-out set consisting of annotations from two domain experts. The system's performance was measured

by the area under the precision-recall curve (AUPRC) and the area under the receiver operating characteristics curve (AUROC). It is important to mention that our deployed system allows choosing a user-defined threshold on the predicted probabilities, because users have different requirements regarding precision and recall (see Sect. 5.2). Hence, we did not optimize the choice of a threshold, but provide accuracy, precision, and recall for a threshold of 0.6 for illustration.

5.1 Cross-Validation

Cross-validation was based on the dataset consisting of 2,581 cardiac CT scans presented in Sect. 3. It contained 1,733 "good" ECGs, 501 corrupted with global noise, and 391 with localized noise. Note that many traces were contaminated with multiple noise patterns of both categories. Experimental results with respect to the global, localized, and unified model are summarized in Table 1.

The results demonstrate that even individual models achieve high performance scores, with a negligible difference in AUPRC and AUROC between RF and SVM, but a slight advantage for RF with respect to precision. We achieved an additional improvement in precision and recall when combining several models into an ensemble. Although the improvement may seem minor, it becomes relevant when considering >10,000 traces. In production, the cost of not identifying a problem, i.e., a false negative, is generally higher than the cost of a false positive. Moreover, our primary objective is to assist technicians in identifying anomalous cardiac CT scanners and not individual ECGs, which results in Table 1 show. Therefore, multiple corrupted ECGs obtained from the same device need to be identified before action is taken, which justifies trading a higher recall for a lower precision – a corrupted ECG should not be missed. We allow the user to individually adjust the threshold, because the trade-off between precision and recall is often situational.

5.2 External Validation

We deployed our system at Siemens Healthcare, where it is used to automatically analyze previously unobserved ECG traces in a real world setting. Our ensemble

Table 1. Cross-validation results for global noise patterns, localized noise patterns, and both types of noise patterns (All) as defined in Sect. 4.1. Accuracy, precision, and recall were computed at a threshold of 0.6.

Metric	Global(RF)	Localized(RF)	All(SVM)	All(RF)	All(Ensemble)
mean AUROC	0.998	0.996	0.996	0.997	0.997
mean AUPRC	0.997	0.989	0.993	0.994	0.996
mean accuracy	0.990	0.981	0.973	0.978	0.983
mean precision	0.990	0.963	0.964	0.979	0.985
mean recall	0.970	0.934	0.952	0.954	0.964

of binary classification models was trained on all 2,581 ECG traces before deployment and processed 150 ECG traces during our evaluation period. Two domain experts independently analyzed these traces manually and assigned each trace to one of 5 categories: (1) perfect (no artifact), (2) good (only very minor artifacts), (3) corrupted (considerable amount of artifacts), (4) strongly corrupted, and (5) extremely corrupted. Overall, the inter-expert agreement was high, as indicated by Kendall's coefficient of concordance ($W = 0.938$, $P < 0.001$, corrected for ties) [12]. Most disagreements (24 of 53; 45.3 %) were due to traces of the 3rd category being assigned to the 2nd (15) or 4th category (9) instead, which indicates that it is difficult, even for experts, to draw a sharp line between clean and corrupted ECGs (see Table 2). We evaluated our system based on AUPRC, AUROC, and Kendall's coefficient of concordance [12], which measures the degree of agreement between our system and the experts on the five-level Likert-scale.

First, we treated categories 3, 4 and 5 as positive class and the remainder as negative class to allow comparison to our cross-validation results. We obtained an AUPRC and AUROC score of 0.875 and 0.898 with respect to expert 1 and 0.905 and 0.942 with respect to expert 2. Although performance scores dropped compared to our cross-validation experiment, it is noteworthy that the expert, who annotated the training set, did not participate in annotating the hold-out set. Thus, corner cases between "good" and corrupted signals are likely biased. The AUROC score still indicates a highly discriminatory model ($\gtrapprox 0.9$) and the drop in AURPC can be attributed to a decrease in precision. Note that expert 1 assigned more ECGs to the positive class (cat. 3-5) than expert 2, thus only 80 % (56/70) of positive annotations of expert 1 match that of expert 2. In contrast, 93 % (56/60) of positive annotations of expert 2 match that of expert 1 (cf. Table 2). Consequently, we would expect that the AUPRC, or average precision, of our system would be around 0.9 at best. We obtained AUPRC scores of 0.875 and 0.905, indicating a highly discriminatory model.

When considering in which range our predicted probabilities fell, we noticed that predicted probabilities of ECGs belonging to categories 2 and 3 were inconsistent. Table 3 shows confusion matrices obtained after dividing predicted probabilities into 5 equally spaced bins ($[0; 0.2[, [0.2; 0.4[, \ldots$). The table reveals that ECGs of categories 2 and 3 have been assigned probabilities in the whole interval $[0; 1]$, which results in a low precision. The recall remains high when disregarding

Table 2. Confusion matrix illustrating inter-expert agreement.

		Expert 1				
		1	**2**	**3**	**4**	**5**
Expert 2	**1**	32	21			
	2		23	11	3	
	3		4	23	6	
	4			3	12	
	5				5	7

Table 3. Confusion matrices demonstrating results of external validation.

		Expert 1					Expert 2				
		1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
Predicted	**1**	27	32	6	1		45	19	2		
	2	3	5	7			5	5	5		
	3		5	4				6	3		
	4		2	3	5			5	3	1	1
	5	2	4	17	20	7	3	2	20	14	11

category 3: the system recognizes 59 out of 60 ECGs of categories 4 and 5 by predicting a probability above 0.6, whereas the recall drops to 0.838 (109/130) when including category 3. At the same time the precision increases from merely 0.492 for categories 4 and 5, to 0.790 for categories 3-5 due to less false positives. We concluded that predicted probabilities are not well calibrated, because very high and very low probabilities are over-represented. In fact, this is a problem for many machine learning methods, which can perform well by means of standard metrics for classification, but yield poorly calibrated probabilistic scores, or vice versa [7]. Alternate learning regimes, such as ordinal regression and learning-to-rank, could remedy this problem. However, in contrast to classification, richer annotations are required, which places more burden on human annotators and makes obtaining labels prohibitively costly in our case.

Next, we compared the model's predictions to the five-level Likert scale, which resulted in Kendall's coefficient of concordance of 0.863 ($P < 0.001$) based on the two expert annotations and the predicted probabilities (corrected for ties). The results demonstrate that most predictions were concordant with the experts' annotations (87.5 % and 82.8 %, excluding ties), thus the agreement between predicted probabilities and expert annotations is substantial.

Although results of the external validation suggest a less discriminatory system, compared to our cross-validation results, the overall performance of the system is still high. Moreover, we allow the user to individually adjust the threshold to identify only severe cases (categories 4 and 5) with very high recall but moderate precision, or all cases (categories 3–5), which increases precision. Overall, the external validation confirmed the practical applicability of our system. Most importantly, the automated analysis operates at a speed that allows processing over thousand ECGs per hour (single-threaded), compared to a few hundred per day of a human analyst.

6 Conclusion

The main goal of this work was to develop a system to automatically detect various non-biological artifacts and noise patterns in ECG signals acquired during cardiac CT. We adapted a discord discovery technique for detecting the most abnormal heartbeats and applied robust PCA for a more precise localization of non-biological anomalies. As a result, we produced a feature set that captures differentiating properties of various global and local noise patterns and used it to train an ensemble of classification models. We validated our system internally via cross-validation and externally in a real world setting. The results demonstrate that our system is highly discriminatory and allows processing thousands of ECGs with minimal human interaction. In the future, we would like to improve our model with regard to calibration, such that predicted scores accurately reflect the true severity of an artifact. Our system is currently deployed at Siemens Healthcare, where it continuously analyzes cardiac CT scans collected from various medical centers. The ultimate benefit of our work can be determined retrospectively as time passes, based on the overall reduction of reported problems and the time needed to resolve them.

Acknowledgments. This work was supported by Siemens Healthcare GmbH.

References

1. Biel, L., Pettersson, O., Philipson, L., Wide, P.: ECG analysis: a new approach in human identification. IEEE Trans. Instrum. Meas. **50**(3), 808–812 (2001)
2. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
3. Candes, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? J. ACM **58**(3) (2011). Article Number 11
4. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
5. De Chazal, P., Dwyer, M.O., Reilly, R.B.: Automatic classification of heartbeats using ECG morphology and heartbeat interval features. IEEE Trans. Biomed. Eng. **51**(7), 1196–1206 (2004)
6. Donoho, D.L.: De-noising by soft-thresholding. IEEE Trans. Inf. Theor. **41**(3), 613–627 (1995)
7. Esarey, J., Pierce, A.: Assessing fit quality and testing for misspecification in binary-dependent variable models. Polit. Anal. **20**(4), 480–500 (2012)
8. Fix, E., Hodges, J.L.: Discriminatory analysis - nonparametric discrimination: consistency properties. Technical report, USAF School of Aviation Medicine, Randolph Field, TX (1951)
9. Hausleiter, J., Meyer, T.S., Martuscelli, E., Spagnolo, P., Yamamoto, H., Carrascosa, P., Anger, T., Lehmkuhl, L., Alkadhi, H., Martinoff, S., et al.: Image quality and radiation exposure with prospectively ECG-triggered axial scanning for coronary CT angiography: the multicenter, multivendor, randomized PROTECTION-III study. JACC Cardiovasc. Imaging **5**(5), 484–493 (2012)
10. Hsieh, J., Londt, J., Vass, M., Li, J., Tang, X., Okerlund, D.: Step-and-shoot data acquisition and reconstruction for cardiac x-ray computed tomography. Med. Phys. **33**(11), 4236–4248 (2006)
11. Karpagachelvi, S., Arthanari, M., Sivakumar, M.: ECG feature extraction techniques - a survey approach. Int. J. Comput. Sci. Inf. Secur. **8**(1), 76–80 (2010)
12. Kendall, M.G., Smith, B.B.: The problem of m rankings. Ann. Math. Stat. **10**(3), 275–287 (1939)
13. Keogh, E., Lin, J., Fum, A.: Hot SAX: Efficiently finding the most unusual time series subsequence. In: 5th IEEE International Conference Data Mining, pp. 226–233. IEEE Computer Society, Washington, DC (2005)
14. Lin, J., Keogh, E., Wei, L., Lonardi, S.: Experiencing SAX: a novel symbolic representation of time series. Data Min. Knowl. Discov. **15**(2), 107–144 (2007)
15. Mallat, S.G.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell. **11**(7), 674–693 (1989)
16. Rosner, B.: Percentage points for a generalized ESD many-outlier procedure. Technometrics **25**(2), 165–172 (1983)
17. Vintsyuk, T.K.: Speech discrimination by dynamic programming. Cybern Syst. Anal. **4**(1), 52–57 (1968)
18. Widrow, B., Glover, J.R., McCool, J.M., Kaunitz, J., Williams, C.S., Hearn, R.H., Zeidler, J.R., Dong, E., Goodlin, R.C.: Adaptive noise cancelling: Principles and applications. Proc. IEEE **63**(12), 1692–1716 (1975)
19. Wolpert, D.: Stacked generalization. Neural Netw. **5**(2), 241–260 (1992)