Contents lists available at ScienceDirect

Smart Health

journal homepage – www.elsevier.com/locate/smhl

ML algorithms to estimate data reliability metric of ECG from inter-patient data for trustable AI-based cardiac monitors

Mst Moriom R. Momota* and Bashir I. Morshed

Department of Computer Science, Texas Tech University, Lubbock, TX 79409, USA

ARTICLE INFO

Keywords:
Cyber-Physical System
Data Reliability Metric
ECG Signal
False Alarm prevention
Noises
Smart Health

ABSTRACT

Improvement in remote healthcare systems has been conducted utilizing cyber-physical systems (CPS), although data reliability is still the biggest challenge in unsupervised settings. Cardiac monitoring remotely using wearable sensors and implementing trustable artificial intelligence (AI) algorithms are one of the most challenging problems in Smart Health (sHealth). Remote cardiac monitoring using wearable electrocardiogram (ECG) devices is prone to noise and artifacts, which can make the data unreliable as false detection, can happen due to the presence of noise. Checking the reliability of data manually is time-consuming and not possible for long-term recorded data from unsupervised settings. To mitigate the data reliability issue in remote cardiac monitoring, we propose a novel Data Reliability Metric (DReM), where we predict the reliability of inter-patient ECG data using machine learning (ML) algorithms from data statistics itself. We predicted DReM on a scale of 0 to 1, where 0 means an extremely noisy signal and 1 means a noise-free signal. In this work, we have used Lasso regression (LR), Support Vector Regression (SVR), Decision Tree Regression (DTR), and Random Forest Regression (RFR) algorithms. To evaluate the model performance R2, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) was used. Decision tree regression was able to predict the DReM with a high correlation ($R^2 = 1$) and low prediction error. We tested our model using different patients' data and observed that DTR performed best in all cases. Our proposed data reliability approach can be beneficial in cardiac disease detection by identifying false predictions from unreliable data.

1. Introduction

In recent years, cyber-physical system (CPS) is being recognized as emerging technology and has received considerable attention since the US National Science Foundation (NSF) designated it as a priority research area in 2008 (Haque et al., 2014). The concept of CPS is based on control, computation, and communication to connect the virtual world with the physical world. In smart healthcare system Patients' physical worlds, medical devices, and equipment are all part of CPS, as both are externally monitored and controlled medical treatments, which are linked to the cyber world via communication networks for information exchange and transfer of physiological data, which is analyzed for feedback and control signals (Verma, 2022). Although Smart Healthcare quality has improved by using CPS, still there are some challenges presents among them reliability is the most important metric. Unreliable data in automated CPS will end up with a high rate of Type II errors and false alarms, which diminishes trust in AI. In some cases, it might be critical for intensive care unit patients.

* Corresponding author.

E-mail address: mmomota@ttu.edu (M.M.R. Momota). bmorshed@ttu.edu (B.I. Morshed)

For cardiac monitoring, electrocardiography is the standard procedure to evaluate cardiac activity by producing an electrocardiogram (ECG/EKG) in a graph of voltage versus time (Kaplan et al., 2018). Although clinically used devices are standard procedures, with the advancement in sensing technology, data fusion, and networking, wearable devices are able to collect the real-time data of the users to monitor their health status from outside of the hospital (Bansal and Joshi, 2018; Zheng et al., 2014). Although almost all the wearable ECG monitoring devices used AI algorithms to predict abnormalities such as Arrhythmia, atrial fibrillation, etc with high accuracy, these devices are prone to noise and artifacts due to weak skin-electrode contact, respiration, and the body movement of users (Satija et al., 2018; Nizami et al., 2013). A study showed that corrupted ECG data with noise and artifacts influence the reliability of the AI algorithm's performance for predicting cardiac arrhythmias and a false alarm was created (Borowski et al., 2011).

Different approaches have been proposed to reduce the false alarm among them ECG denoising based strategy and signal quality index (SQI) based strategy to assess the clinical acceptability of the recorded ECG signals is the most popular (Satija et al., 2018; Orphanidou and Drobnjak, 2017). Although denoising-based strategy can easily remove high and low-frequency noises from the ECG signal using different filtering techniques such as high pass filter, low pass filter, moving average filter, and so on, removing muscle artifacts is still challenging as it mimics the morphology of the ECG waveform. Other limitations of the denoising technique were reported for instance distortion of amplitude reduction, shape alteration, and width widening (Satija et al., 2016). In some cases, normal heartbeats are misclassified as abnormal heartbeats due to the distortion of local waves such as P wave, QRS complex, and T wave. To evaluate the clinical acceptability of the recorded ECG signals, SQI based strategy is used. Using the SQI strategy an algorithm was proposed to classify true and false alarms of arrhythmia in the intensive care unit with an average of 75.54% accuracy using a random forest classifier (Eerikainen et al., 2015). A generalized approach was proposed for ECG and PPG data reliability estimation and achieved 99.7% cross-validation accuracy using the Random Forest classifier (Zaman and Morshed, 2020).

In this work, we proposed a novel approach to estimate the reliability of inter-patient ECG data on a scale of 0 to 1, where 0 means data is not reliable at all and highly contaminated with noise, and 1 means data is highly reliable with very little or no noise. Checking data reliability manually is not feasible for a continuous monitoring system. We have used the machine learning (ML) technique to estimate this reliability metric from the signal statistics. In this approach, we have first extracted both time-series features (statistical, spectral, and temporal) and morphological features from the noise-contaminated ECG data. To minimize the computational cost, we used the least number of top-ranked features required for our analysis. Then we trained our regression model and determined the best-performed model on test sets. Our proposed automated method can be useful in remote cardiac monitoring to identify false alarms due to the less reliable data in ECG, thus improving the trust in AI-based cardiac monitoring.

2. Methodology

2.1. System Overview

The proposed "Data Reliability Metric" (DReM) is a measure (between 0.0 to 1.0) that indicates reliability of data collected from unsupervised settings. Lower value of DReM represents less reliable data, whereas a higher value represents more reliable data. DReM can be computed by combining sensor specific functions (application domain indices) with CPS generalizable functions (CPS domain indices). For ECG/EKG signals, application domain indices can be time interval features (P-P, Q-Q, R-R, S-S, T-T, T-T′, P-Q, P-R, P-S, P-T, Q-R, Q-S, Q-T, R-S, S-T) and amplitude features (P, Q, R, S, T, PQ, PR, PS, PT, QR, PR, PS, PT, QR, QS, QT, QT′, RS, ST). CPS domain indices can include statistics (e.g., kurtosis, skewness), feature vector progression over time and across nodes, coherence of Principle Component Analysis (PCA) and low Intra-Class Correlation (ICC), and heuristic match with past records.

The concept of DReM and how it can be beneficial in decision-making before sending feedback to the users is depicted in Figure 1. In a smart healthcare system for remote cardiac monitoring real-time ECG data is collected through wearable ECG monitoring devices. Those data are processed to find out abnormalities for early cardiac disease detection. As wearable devices are vulnerable to noise and artifacts for automated abnormalities detection methods can be affected and end up with an inaccurate prediction. To avoid false alarms in abnormality detection, we proposed a novel approach for automated data reliability estimation to ensure precise decision-making.

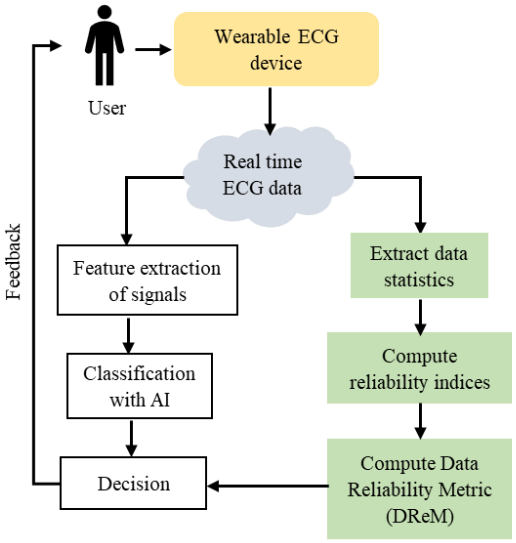

Fig. 1. Concept of Data Reliability Metric (DReM) for a CPS-based ECG wearable device for early cardiac abnormalities detection.

Figure 2 illustrates the flowchart of our work for computing DReM. Firstly, we prepared the dataset using clean ECG datasets and the noisy database. After that signal-specific and signal-independent features are extracted and then selected important features that are required for our work. Then we trained the model with training data and then test the model with different data sets and evaluate the performance of the model using R^2, Mean Absolute Error, and Root Mean Square Error scores. Finally, we computed DReM on a regression scale of 0 to 1.

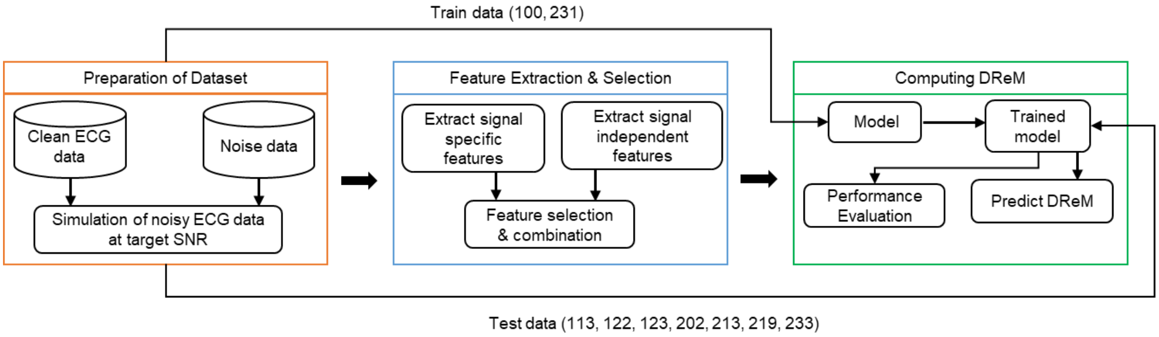
Train data (100, 231)
Preparation of Dataset
Clean ECG data
Noise data
Simulation of noisy ECG data at target SNR
Feature Extraction & Selection
Extract signal specific features
Extract signal independent features
Feature selection & combination
Computing DReM
Model
Trained model
Performance Evaluation
Predict DReM
Test data (113, 122, 123, 202, 213, 219, 233)

Fig. 2. Flowchart of ML method for data reliability metric (DReM) computation from noisy ECG signals.

2.2. ECG Dataset

MIT-BIH Arrhythmia Database and MIT-BIH Noise Stress Test Database are used in this study (Moody and Mark, 2001; Moody et. al., 1984). These databases are collected from PhysioNet, which is a research source for complex physiological signals (Goldberger et al., 2000). MIT-BIH Arrhythmia Database records are 30 minutes long with a sampling frequency of 360 Hz. For noise source, we used MIT BIH Noise Stress Test (NST) database, which contains baseline wander (BW), muscle (EMG) artifact (MA), and electrode motion artifact (EM) noises, which are the most common noise present in ECG signal. BW is a low-frequency noise (frequency band is below 1 Hz) that occurred due to the subject's respiration movements; lack of skin-electrode contact might be another reason for BW. MA is a high-frequency noise (frequency ranges between 20-1000 Hz) that happened during rapid body movement, which is also known as EMG noise as it happened due to muscle electrical activity. EM is considered the most troublesome, as it can mimic the appearance of ectopic beats and occurred due to electrode motion (frequency ranges from 1 Hz to 15Hz).

2.3. Noisy Signal Creation

From the MIT-BIH noise stress database BW, EM, and MA are mixed with a clean ECG signal with all possible combinations of these three noises to generate a noisy signal. Figure 3 shows the process of our noisy signal generation at the target SNR. For noisy signal generation at the target SNR, we have followed the NST open-source documentation. The following equation is used in the noise generation algorithm to calculate the target SNR.

$$SNR = 10\log_{10}\frac{S}{N \times a^2} \qquad (1)$$

Where S and N are the signal power and the noise power respectively. The scaling factor "a", is used along with the SNR equation to calculate the target SNR. For this study, all possible combinations of the noisy signal are generated at SNR levels -20 dB to 20 dB with step size 1.

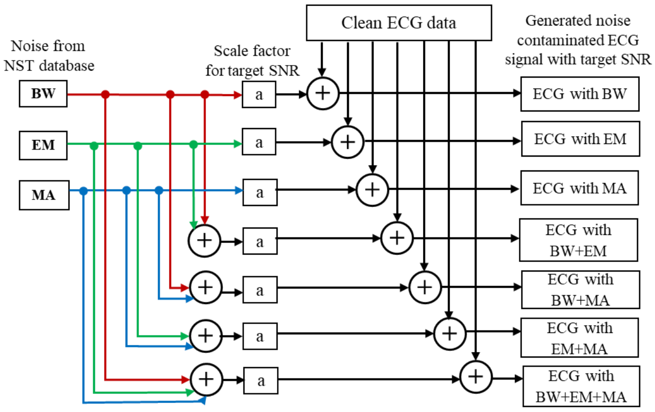

Fig. 3. The procedure of generating all possible noisy ECG signals using BW, EM, and MA at the target SNR value.

2.4. Feature Extraction and Selection

Signal independent features and signal-specific features also known as morphological features are extracted in our work. We took 10 seconds of data and extracted features and shifted until it finishes the 30 min duration. For signal-independent feature extraction, we used the Time series feature extraction library (TSFEL) in python, which extracts features in statistical, temporal, and spectral domains (Barandas et al., 2020). 391-time-series features are extracted using the TSFEL library. Some features are correlated with each other and not all the features are important for our work. As lower feature numbers can reduce the computation cost, which is an important factor for low-power wearable devices, we selected the minimum number of required features based on our requirement. To rank our extracted features based on their importance, we have used the embedded feature selection technique with Lasso Regression (LR) learning algorithm. As the embedded feature selection method is a combination of the filter method and wrapper method it is more accurate than other feature selection techniques. We have selected the top 10 time-series features based on feature importance. For morphological feature extraction, PySiology python package was used, which extracted 10 signal-specific features (Gabrieli et al., 2020). Table 1 summarizes the extracted and selected top-ranked features.

Table 1

Overview of signal independent, signal specific, and selected top 10 features

Feature type	Feature list
Signal-independent features (time-series features) extracted using TSFEL library	391 features are extracted in statistical, temporal, and spectral domains.
Signal-specific feature (morphological feature) extracted using PySiology library	Inter-beat intervals (IBIs), Beat per minute (BPM), Standard deviation of NN intervals (SDNN), Standard deviation of the differences between adjacent IBIs (SDSD), Square root of the mean of the sum of the squares of the differences between adjacent IBIs (RMSSD), Average number of consecutive normal sinus (NN) intervals that change more than 50 milliseconds (Pnn50), Number of pairs of successive NNs that differ by more than 20 ms (Pnn20), Low-frequency power (LF), High-frequency power (HF) and Very low frequency (VLF).
Selected top 10 features from time-series feature using embedded feature selection technique	Kurtosis, Skewness, Spectral entropy, Spectral centroid, Median frequency, Median absolute difference, Power bandwidth, Histogram, Root mean square, Linear prediction cepstral coefficients

2.5. DReM Computation

In this work, we have used Least Absolute Shrinkage and Selection Operator (Lasso) Regression (LR), Support Vector Regression (SVR), Random Forest Regression (RFR), and Decision Tree Regression (DTR) algorithms to predict the DReM. The grid search technique is used to come up with a combination of optimum parameters for these models. The optimum hyperparameters of these regression models are illustrated in Table 2. We have trained the model with our training data sets (100 and 231) with the labeling of actual data reliability metrics which were calculated from SNR. Model performance is evaluated with R^2, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). R^2 score is used to measure the variance in predictions made by the model, and the score of R^2 is ranging from 0 to 1. $R^2 = 1$ represents when modeled values exactly match the observed values. MAE measures the amount of error in prediction by calculating the difference between the predicted value and true value. RMSE is the standard deviation of the prediction errors, which measures the spread out of the prediction errors. Using test data with a trained model we have validated the model performance and computed the DReM.

Table 2

Optimum parameter of the models

Model	Combination of hyperparameter
LR	$\alpha = 1.0$; tol = 0.0001; selection = cyclic
SVR	C=1.0; tol=0.0001; kernel=poly; D=3; ε=0.001
RFR	Number of trees = 200; maximum depth = 80; maximum leaf nodes = 20
DTR	Maximum depth = 6; maximum leaf nodes = 20; minimum sample leaf = 40

3. Results

3.1. Simulated Noisy Signal

Figure 4 (a-g) shows the example of our simulated noisy ECG signal with all 7 combinations of noise and artifacts of BW, EM, and MA. The noisy signal was simulated at -20 dB to 20 dB SNR with step 1. In figure 4 noise-contaminated ECG signals are portrayed at -20 dB, 0 dB, and 20 dB SNR.

3.2. Performance Evaluation

Lasso Regression (LR), Support Vector Regression (SVR), Random Forest Regression (RFR), and Decision Tree Regression (DTR) model's performance are evaluated for test data 113, 122, 123, 202, 213, 219, and 233. Our selected top 10 time-series features and 10 signal-specific features are used separately and combined for evaluating the model performance and computing the DReM of these test data sets. Table 3 shows the average RMSE, and MAE scores of all test data, while figure 5 illustrates the average R^2 values of SVR, LR, RFR, and DTR for all test data using the top 10 time-series features, all extracted signal-specific features, and a combination of top 10 time-series and 10 signal-specific features. LR, RFR, and DTR algorithms performed with lower prediction error (low MAE and low

RMSE) than SVR in all cases. Figure 5 shows that SVR achieved an average R^2 value of 0.96 when we used only the top 10 time-series features. For 10 signal-specific features, SVR performance gets worse which affects the model performance during combined features use. LR, RFR, and DTR showed promising performance for all types of features with a high R^2 score, which represents a high correlation between predicted DReM and actual DReM. In this work the ground truth of our work is based on the SNR.

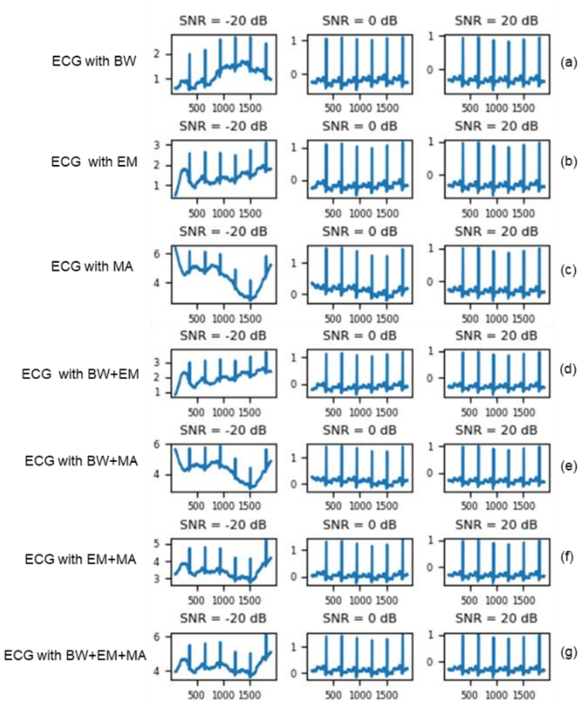

Fig. 4. Example of our simulated noise contaminated ECG signal at -20 dB, 0 dB, and 20 dB SNR. ECG signal with (a) BW, (b) EM, (c) MA, (d) BW + EM, (e) BW + MA, (f) EM + MA, and (g) BW+ EM + MA.

Table 3

Average test performance of all test data using time series, signal specific, and a combination of these two features

Features	Algorithms	RMSE	MAE
Top 10 time series features	SVR	6.04e-02	5.02e-02
	LR	1.03e-02	8.37e-03
	RFR	1.26e-05	6.74e-07
	DTR	**1.18e-15**	**6.95e-16**
Extracted 10 signal specific features	SVR	1.97e-01	1.39e-01
	LR	8.24e-03	6.84e-03
	RFR	1.47e-05	8.15e-07
	DTR	**1.31e-15**	**8.29e-16**
Top 10 time series features + 10 signal specific features	SVR	2.36e-01	1.6e-01
	LR	1.15e-02	9.3e-03
	RFR	1.09e-05	5.34e-07
	DTR	**1.26e-15**	**7.68e-16**

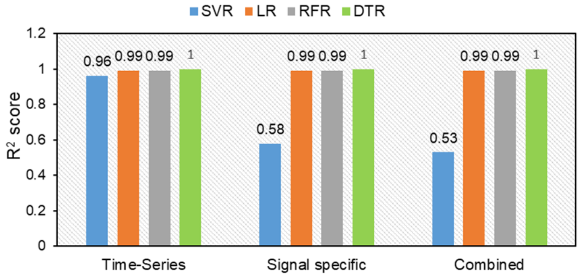

Fig. 5. Average R^2 of regression models for all test data using top 10 time-series features and 10 signal-specific features separately and combined.

Table 4 summarizes the performance of our best-performed models LR, RFR, and DTR models with different test datasets. Lasso regression and random forest regression achieved an R^2 value of 0.99. Decision tree regression is our best-performed model for all data sets with the highest R^2 value (R^2 = 1). The mean absolute error and root mean square error was very low for all test data while we used decision tree regression and random forest regression algorithm. We can say from Table 4 that among all three algorithms, Decision Tree Regression (DTR) is the best-performed model.

Table 4

Performance of ML models of all test data sets using top-ranked time series and signal-specific features

Test Data	Algorithms	R^2	RMSE	MAE
113	LR	0.9988	9.9e-03	8.0e-03
	RFR	0.9999	7.01e-06	1.97e-07
	DTR	**1**	**1.3e-15**	**8.01e-16**
122	LR	0.9982	1.22e-02	9.7e-03
	RFR	0.9999	1.21e-05	5.9e-07
	DTR	**1**	**1.19e-15**	**6.94e-16**
123	LR	0.9977	1.4e-02	1.12e-02
	RFR	0.9999	1.22e-05	5.91e-07
	DTR	**1**	**1.4e-15**	**9.1e-16**
202	LR	0.9989	9.4e-03	7.7e-03
	RFR	0.9999	1.71e-05	7.86e-07
	DTR	**1**	**1.27e-15**	**7.76e-16**
213	LR	0.9981	1.28e-02	1.04e-02
	RFR	0.9999	1.4e-05	7.87e-07
	DTR	**1**	**1.08e-15**	**6.09e-16**
219	LR	0.9979	1.33e-02	1.08e-02
	RFR	0.9999	1.4e-05	7.87e-07
	DTR	**1**	**1.31e-15**	**7.86e-16**
233	LR	0.9982	1.22e-02	1.0e-02
	RFR	0.9999	1.57e-05	9.83e-07
	DTR	**1**	**1.29e-15**	**7.98e-16**

3.3. Predicting DReM

We have calculated the actual DReM from the ground truth, SNR. -20 dB to 20 dB SNR are linearly normalized to a scale of 0 to 1 to calculate the actual DReM from the SNR. Figure 6 illustrates the prediction of the data reliability metric on test data 113 using SVR, LR, RFR, and DTR using a combination of our selected top 10 time series and extracted 10 signal-specific features. LR, RFR and DTR are highly correlated with the predicted value. We have predicted DReM with other test data and in all cases, the best model was Decision Tree Regression.

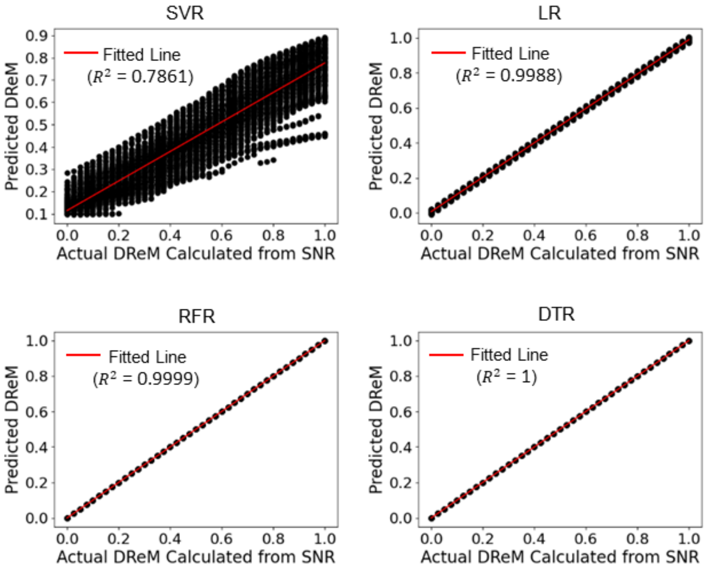

Fig. 6. Predicting DReM of test data 113 using LR, SVR, RFR, and DTR algorithms.

4. Conclusions

Wearable ECG monitoring devices are vulnerable to noise and artifacts. ECG data corrupted with noise can hamper the abnormality detection in cardiac monitoring. Thus, before taking any decision only relying on AI algorithms might be harmful to patients in some cases life-threatening. Therefore, checking the reliability of data is the most important step before making a decision. But checking the reliability of data is time-consuming and not always possible if the data is too long. In this study, we have proposed a novel automated data reliability metric (DReM) to predict the reliability of data on a regression scale (between 0 to 1, where 0 means data is not reliable, and 1 means data is reliable). We have used both signal-independent and signal-specific features and used lasso, support vector, random forest, and decision tree regression ML algorithms for training and testing data. For evaluating model performance R^2, RMSE, and MAE are used as analysis indices. LR, RFR, and DTR algorithms performed well for both signal-independent and signal-specific features with high R^2 scores and low prediction error. But SVR performance gets worse with the morphological feature. Decision tree regression is the best-performed model and for all test data sets, it can predict DReM with the highest R^2 score ($R^2 = 1$). Our proposed automated data reliability method can be useful for the wearable cardiac monitoring device to increase trust in AI-based cardiac monitoring algorithms.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1932281.

References

Haque, S.A., Aziz, S.M. and Rahman, M. (2014) ‘Review of cyber-physical system in healthcare’, *International Journal of Distributed Sensor Networks*.

Verma, R. (2022) ‘Smart City Healthcare Cyber Physical System: Characteristics, Technologies and Challenges’, *Wireless Personal Communications*, 122(2), pp. 1413–1433.

Kaplan, J. A., Cronin, B. and Maus, T. (2018) ‘Kaplan’s Essential Cardiac Anesthesia for Cardiac Surgery by Joel (second edition)’, *Elsevier.* ISBN: 978-0-323-49798-5.

Bansal, A. and Joshi, R. (2018) ‘Portable out-of-hospital electrocardiography: A review of current technologies’, *Journal of Arrhythmia*, 34(2), pp. 129–138.

Zheng, Y.-L. *et al.* (2014) ‘Unobtrusive Sensing and Wearable Devices for Health Informatics’, *IEEE Transactions on Biomedical Engineering*, 61(5), pp. 1538–1554.

Satija, U., Ramkumar, B. and Manikandan, M.S. (2018) ‘A Review of Signal Processing Techniques for Electrocardiogram Signal Quality Assessment’, *IEEE Reviews in Biomedical Engineering*, 11(c), pp. 36–52.

Nizami, S., Green, J.R. and McGregor, C. (2013) ‘Implementation of Artifact Detection in Critical Care: A Methodological Review’, *IEEE Reviews in Biomedical Engineering*, 6, pp. 127–142.

Borowski, M. *et al.* (2011) ‘Medical device alarms’, *Biomed Tech (Berl).* 56(2), pp.73-83.

Orphanidou, C. and Drobnjak, I. (2017) ‘Quality Assessment of Ambulatory ECG Using Wavelet Entropy of the HRV Signal’, *IEEE Journal of Biomedical and Health Informatics*, 21(5), pp. 1216–1223.

Satija, U., Ramkumar, B. and Manikandan, M.S. (2016) ‘A unified sparse signal decomposition and reconstruction framework for elimination of muscle artifacts from ECG signal’, in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 779–783.

Eerikainen, L.M. *et al.* (2015) ‘Decreasing the false alarm rate of arrhythmias in intensive care using a machine learning approach’, in *2015 Computing in Cardiology Conference (CinC)*. IEEE, pp. 293–296.

Zaman, M.S. and Morshed, B.I. (2020) ‘Estimating Reliability of Signal Quality of Physiological Data from Data Statistics Itself for Real-time Wearables’, in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, pp. 5967–5970.

Moody, G.B. and Mark, R.G. (2001) ‘The impact of the MIT-BIH arrhythmia database’, IEEE Engineering in Medicine and Biology Magazine, 20(3), pp. 45–50.

Moody, G.B. Muldrow, We. and Mark, R (1984). The mit-bih noise stress test database.

Goldberger, A.L. *et al.* (2000), ‘PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals’, Circulation. 101 (23), pp. e215–e220.

Barandas, M. *et al.* (2020) ‘TSFEL: Time Series Feature Extraction Library’, *SoftwareX*, 11, p. 100456.

Gabrieli, G., Azhari, A. and Esposito, G. (2020) ‘PySiology: A Python Package for Physiological Feature Extraction’, in *Smart Innovation, Systems and Technologies*, pp. 395–402.