Outlier detection in ECG

Omar Atamny[1,2], Ardan Saguner[3], Roger Abaecherli[4], Ender Konukoglu[1]

[1] ETH Zurich, Zurich, Switzerland
[2] Technical University of Munich, Munich, Germany
[3] University Hospital of Zurich, Zurich, Switzerland
[4] Lucerne University of Applied Sciences and Arts, Lucerne, Switzerland

Abstract

Automatic abnormal ECG detection algorithms are crucial for treating heart health problems and saving lives. This study's objective is to verify whether using unsupervised learning methods and more specifically probabilistic and deep learning models such as autoencoders, variational autoencoders (VAE), diffusion Models, partial shift variational autoencoder (our implementation of half VAE half prediction model), normalizing flows and Gaussian mixture models to detect outliers in ECG data is possible. An outlier for our case is an abnormal ECG signal, a one belonging mostly to a sick person while the normal case is that of a normal healthy person. The results have shown that the models distinguish between normal and abnormal data to a specific degree, with the VAE achieving an area under the curve (AUC) of the receiver operating characteristic curve score of 0.85 on the publicly available PTB-XL dataset and 0.83 on the publicly available CPSC dataset. Moreover, the VAE achieved an AUC of 0.89, 0.80 and 0.81 when distinguishing between normal and conduction disturbance, myocardial infarction, and ST/T Change respectively. This indicates that a VAE when optimized itself and fed with more proper data may be able to be used in medical applications.

1. Introduction

Cardiac problems are the number one reason for death globally causing 17.9 million deaths globally annually [1]. Electrocardiography (ECG) is a safe, noninvasive and easy to do diagnostic test to detect heart disease, it is the most frequent cardiovascular test with nearly 200 million ECGs recorded annually in the globe, it is crucial for the evidence-based management of cardiovascular conditions [2]. For the ECG case, be able to classify abnormalities by their deviations from the normative distribution through calculating the error between the input and reconstruction. In this work, we are concerned with answering the question whether we can automatically detect abnormalities (outliers) in ECG data in an unsupervised way.

Previously, authors used machine leaning methods to classify ECG rhythms and beats into healthy or not or into more classes using probabilistic and deep learning models such as convolutional neural networks, adversarial autoencoders and improved AnoGANs. The works of [3] and [4] were unsupervised methods trained on normative data to detect abnormal beats while that of [5] was a supervised learning method which had the normal and abnormal data as well as the labels for training to classify abnormal rhythms. This work contribution is in using unsupervised learning methods and more specifically probabilistic models for abnormal rhythm detection to contribute to the Physionet challenge 2020 [6].

2. Methods

2.1. Autoencoders (AEs)

AEs map an input signal to a, usually, lower dimensional latent representation with the encoder part, and then reconstruct the input signal from that representation through the decoder part with the intention that the output is as much close to the input as possible. The lower dimensionality is deliberately creating a bottleneck, so that the network has to compress crucial information about the signal in that representation that allows reconstructing the signal from it back [7]. We trained an AE with the normative data and detected abnormalities by reconstruction error, assuming data falling out of training distribution would incur a large reconstruction error as these reconstructions would be shifted more toward the normal data and thus will have a higher difference between input and output.

2.2. Variational autoencoders (VAEs)

VAEs have a similar architecture to AEs, however, the latent space vector is sampled stochastically from the latent distribution [8] and thereafter the decoder reconstructs

the output. The encoder part, encodes the input data into a lower dimensional latent space until it outputs a set of means and variances. Different than AEs, VAEs optimize the evidence lower-bound to the data likelihood, which consists of a reconstruction term, as AEs, and a KL-divergence term; together these terms allow VAEs to approximate high-dimensional distributions. Like the AE, we trained a VAE using normative data and detected abnormalities through the reconstruction error as the outlier data was reconstructed more toward the normative data and had higher error values.

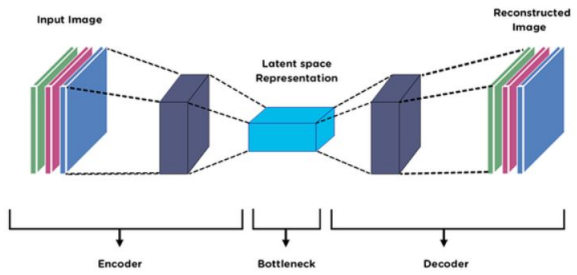

Figure 1. The structure of an autoencoder [9]

2.3. Diffusion models (DMs)

DMs are a type of generative models using neural networks. They are split into two parts, the former adds noise to the image and the later tries to denoise the image back to the original input through approximating the high-dimensional data distribution through this process [10]. The first process is called forward diffusion while the latter is referred to as reverse diffusion process. Both parts are successive steps of nosing or denoising thus DMs could be seen as a chain, however, both parts could be made through a single computation. We trained this model also with normative data. Samples were first fed to the forward diffusion process and reconstructed with the reverse one. Reconstruction error is computed between the original and the reconstructed where higher scores mean an anomaly as it is not reconstructed as good as the normal data.

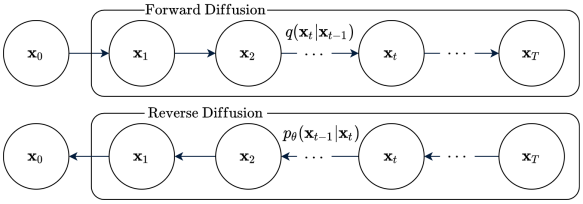

Figure 2. The process of a diffusion model [11]

2.4. Normalizing flow models (NFs)

NFs are a subset of machine learning that uses the change of variable principle to change the input into some other distribution. The functions used are invertible and have easily computable Jacobean matrices [12]. This allows maximizing directly the log-likelihood of data samples without relying on approximations or lower bounds. This method was trained on normative data and classified anomalies in their algorithmic high loss value of the predictions as it is assumed the algorithm would behave worse for data whose distribution was unknown to it.

2.5. Partial shift variational autoencoder

Our model, called PSVAE, takes part of the original signal as the input instead of the complete signal, contrary to normal VAE, and predicts the complete signal back from the latent representation drawn from the part and an operator that specifies where the signal was cut. Like the VAE, the sampling of the data point after the down-sampling and reaching the latent space is done as shown in the following equation with μ as the means and Var as the variance: $Z = \mu + noise * \mathrm{Var}$. This method classified the signals in the same way as the AE and the VAE.

2.6. Gaussian mixture models (GMMs)

GMMs assume a function, or a distribution is mixture of Gaussian distributions, each has its own mean and variance parameters, in addition to that, each Gaussian distribution is weighted by factor such that the sum of all factors is equal too 1 as shown if the equation below [13].

$$\mathrm{X} = \sum w_k \times \mathcal{N}(\mu,\, \sigma^{\in})\,. \\ \sum w_k = 1$$

Same as the normalizing flow method, this model was trained on normal data and classified abnormalities in their higher algorithmic loss values as the abnormal distribution was unkown to the algorithm.

2.7. Data

2.7.1. PTB-XL dataset

The PTB-XL dataset is a large publicly available ECG dataset [14]. This resource contains 21799 12-Lead ECG recording of 18869 patients of different categories and 9517 healthy volunteers as shown in Table 1.

2.7.2. CPSC 2018 dataset

China physiological signal challenge (CPSC) 2018 is China's first physiological signal challenge [15]. It is a

Table 1. PTB-XL category table

Number of records	Description
9517	Normal ECG
5473	Myocardial infarction
5237	ST/T change
4901	Conduction disturbance
2649	Hypertrophy

publicly available dataset containing 12 lead ECG signals of different types and number of recordings as shown in table 2.

Table 2. CPSC 2018 category table

Number of records	Description
918	Normal ECG
1098	Atrial fibrillation
704	First-degree atrioventricular block
207	Left bundle branch block
1695	Right bundle branch block
556	Premature atrial contraction
672	Premature ventricular contraction
825	ST-segment depression
202	ST-segment elevated

2.7.3. Training test split

The training data for all models other than the GMM was 5699 and 700 samples and the testing data was 13901 and 1083 for the PTB-XL and CPSC datasets respectively. For the GMM the training data was 700 and 218, and the testing data was 2000 and 1365 for the PTB-XL and CPSC datasets respectively as it did worse for larger amounts of data. The data was split based on index, where points whose number was higher than the index were test and for lower ones it was for training.

3. Results

The following two tables show the AUC score of the five different methods for the PTB-XL (first table) and CPSC 2018 (second table) datasets. The tables illustrate the scores based on the reconstruction error that specifies if a point is an outlier or not. The error measurement metrics are the mean absolute error (MAE) and the mean squared error (MSE) of the true and output sample, furthermore the algorithmic loss is used for the NF and GMM models. The VAE achieved the best results with AUC scores of 0.88 and 0.84 with the best detection metrics on the PTB-XL and CPSC 2018 datasets, while the overall worst performance was that of the GMM with AUC of 0.69 and 0.70. The other four models lied in between.

Table 3. AUC Table - PTB-XL dataset

Model	AUC MAE	AUC MSE	AUC Loss
Autoencoder	0.72	–	–
VAE	0.84	0.88	–
Diffusion model	0.74	0.76	–
Normalizing flow	–	–	0.73
PSVAE	0.73	0.80	–
GMM	–	–	0.70

Table 4. AUC Table - CPSC 2018 dataset

Model	AUC MAE	AUC MSE	AUC Loss
Autoencoder	0.76	–	–
VAE	0.83	0.65	–
Diffusion model	0.65	0.63	–
Normalizing flow	–	–	0.71
PSVAE	0.74	0.62	–
GMM	–	–	0.69

As the VAE achieved the best results, we further analyzed it's detection performance per abnormality type in the PTB-XL data set. Results are shown in Table 5. The VAE achieved a 0.89 AUC score for conduction disturbance with the MAE as the highest score, while the lowest score was 0.54 for hypertrophy with MSE.

Table 5. AUC table VAE PTB-XL different categories

Abnormal Class	AUC MAE	AUC MSE
Abnormal - mixed	0.85	0.83
Conduction disturbance	0.89	0.85
Hypertrophy	0.60	0.54
Myocardial infarction	0.80	0.77
ST/T change	0.79	0.81

4. Discussion

Unsupervised learning introduces the opportunity to classify anomalous and normal ECG signals from each other with only training on normative data. Comparing the autoencoder models visually, the VAE produced the best reconstruction results, i.e. its outputs are the most similar to its input (not shown here). Autoencoders with the diffusion model is not an apple to apple comparison, however we note that the diffusion model produced the best reconstructions.

Related work can be split majorly into two parts: supervised which different from our work in having the labels

for training and unsupervised is that like our model does not need labels for training.

The work we can fully compare our AUC with to our knowledge is that of Smigiel et al. [5]. The AUC score of the best model in [5] is 0.96 for the binary classifier while our best model achieved 0.88 AUC on the same publicly available dataset (PTB-XL dataset). It is to be noted however that their model is a supervised learning method, which has more data to train on (the labels) and thus should produce better results.

Another work to be noted is the work of Shan et al. [3] which did beat classification into normal or abnormal categories. Their work achieved AUC scores of 0.96 and 0.93 on the MIT-BIH and CMUH datasets. This is higher than our 0.83 to 0.88 AUC score and that of Shin et al. [4] that achieve an AUC OF 0.94 which is also a beat classification problem.

Our work implies that unsupervised learning methods can be trained to differentiate between normal and outlier ECG data in a good to excellent way as it achieving an AUC of over 0.80 and even more for some specific categories like conduction disturbance with AUC of 0.89. Applications for our research and models are medical and non-medical (like using devices with built in ECG sensors) classification of ECG signals.

Our works limitations are: (i) small sample sizes used for training and validation, (ii) using raw signals instead of pre-processed signals, and (iii) the models are not fine-tuned in terms of hyper-parameters.

5. Conclusion

The study has shown that the models and especially the best performing model the VAE with an AUC of 0.85 on the publicly available PTB-XL dataset and 0.88 on the publicly available CPSC dataset could distinguish between normal ECG data and outliers (abnormal data). Furthermore, the VAE has high classification performance for conduction disturbance, myocardial infarction and ST/T change with an AUC of 0.89, 0.80 and 0.81 respectively. While this performance is not enough for use in medical applications, future work could build on our work and enhance it, more specifically, future research could try new methods to score the difference between input and reconstruction, fine-tune the models and expand them, use different prepossessing methods and try the models on new datasets.

References

[1] Cardiovascular diseases world health organization. URL https://www.who.int/health-topics/cardiovascular-diseases. [Online; accessed April 27, 2023].

[2] Reichlin T, et al. Advanced ecg in 2016: is there more than just a tracing? Swiss Med Wkly Apr. 2016;146:w14303.

[3] Shan L, et al. Abnormal ecg detection based on an adversarial autoencoder. Front Physiol Sep. 2022;13:961724.

[4] Shin DH, et al. Decision boundary-based anomaly detection model using improved anogan from ecg data. IEEE Access 2020;8:108664–108674.

[5] Śmigiel S, et al. Ecg signal classification using deep learning techniques based on the ptb-xl dataset. Entropy Apr. 2021;23(9):1121.

[6] Alday EAP, et al. Classification of 12-lead ecgs: the physionet/ computing in cardiology challenge 2020. Physiological Measurement 2020;.

[7] Goodfellow I, et al. Deep Learning. MIT Press, 2016. http://www.deeplearningbook.org.

[8] Kingma DP, Welling M. An Introduction to Variational Autoencoders. 2019.

[9] Birla D. Basics of autoencoders, 2019. URL https://medium.com/@birla.deepak26/autoencoders-76bb49ae6a8f. [Online; accessed April 27, 2023].

[10] Yang L, et al. Diffusion Models: A Comprehensive Survey of Methods and Applications. Association for Computing Machinery, 2023.

[11] Das A. An introduction to diffusion probabilistic models, 2021. URL https://ayandas.me/blog-tut/2021/12/04/diffusion-prob-models.html. [Online; accessed April 27, 2023].

[12] Kobyzev I, et al. Normalizing flows: An introduction and review of current methods. IEEE Transactions on Pattern Analysis and Machine Intelligence May 2020;PP.

[13] Zemel R, Urtasun R. Mixtures of gaussians and em, 2016. Course: CSC 411, University of Toronto, Canada.

[14] Wagner P, et al. Ptb-xl, a large publicly available electrocardiography dataset (version 1.0.3). PhysioNet 2020;.

[15] Liu F, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. Journal of Medical Imaging and Health Informatics 2018;8:1368–1373.

Address for correspondence:

ETH Zuirch, Raemistrasse 101, 8092 Zurich, Switzerland, Kender@vision.ee.ethz.ch

Techincal University of Munich - TUM, Arcisstraße 21, 80333, Munich, Germany, Atamny.1@gmail.com

University Hospital of Zurich, Raemistrasse 100, 8091 Zurich, Switzerland, Ardan.saguner@usz.ch

Lucerne University of Applied Sciences and Arts, Werftestrasse 4, 6002 Luzern, Switzerland, Roger.abaecherli@hslu.ch