2023 19th International Symposium on Medical Information Processing and Analysis (SIPAIM) | 979-8-3503-2523-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/SIPAIM56729.2023.10373478

Generative Data by β-Variational Autoencoders Help Build Stronger Classifiers: ECG Use Case

Yousef Nademi[1], Sunil V Kalmady[1,2,3], Weijie Sun[1], Amir Salimi[1],
Abram Hindle[1], Padma Kaul[2,3], Russell Greiner[1,4]
[1]*Department of Computing Science*, *University of Alberta*, Edmonton, Canada
[2]*Canadian VIGOUR Centre, Department of Medicine*, *University of Alberta*, Edmonton, Canada
[3]*Department of Medicine*, *University of Alberta*, Edmonton, Canada
[4]*Alberta Machine Intelligence Institute*, Edmonton, Canada
nademi@ualberta.ca, *kalmady@ualberta.ca*

***Abstract*—We explore the challenge of learning models that use electrocardiogram (ECG) data to diagnose various cardiovascular diseases. Here, we explore whether classifiers trained on a dataset of real labeled ECGs, augmented with synthetic ECGs, can perform better than ones trained on unaugmented datasets.**

We first used a dataset of ECGs, each labelled with one or more of 15 diagnoses, from 244,077 patients to train an unsupervised β–VAE model, that could generate time series of 12-lead ECG signals for each of the diagnoses. We then used this generative model to generate ECGs with the ST-segment Elevated (STE) abnormality, which we added to the public dataset of ECG abnormalities (n = 6877, over normal (Sinus Rhythm) and 8 different abnormalities) of China Physiological Signal Challenge 2018, and found a learner trained on this extended dataset performed better than one trained on only the original data on the targeted STE label but also enhanced its performance for the classification of 4 other labels.

***Index Terms*—Electrocardiogram, Machine Learning, Variational Autoencoder, Data Generation**

I. Introduction

Cardiovascular diseases (CVDs) are one of the leading causes of mortality in recent decades, where 233.1 per 100,000 people died globally due to CVDs in 2017 [1], [22]. To reduce mortality caused by such diseases, early diagnosis is crucial. A widespread tool currently used for the diagnosis of cardiovascular diseases is Electrocardiogram (ECG). However, detecting cardiac abnormalities through ECG is not easy and currently requires an expert. With the advancement of machine learning in healthcare, many researchers are now exploring ways to learn end-to-end diagnostic models using ECGs [2], [3], [23]. The open-source ECG data from the China Physiological Signal Challenge 2018 (CPSC 2018) has helped researchers to develop various machine-learning models for ECG abnormalities/ diagnosis. However, the prediction performance of these models is not high for all labels. While it is well-known that increasing the training dataset with real ECG data can improve the prediction performance, whether synthetic ECGs can do the same remains to be explored.

Since the ECG is routinely used at point of care, it is one of the most common measurements; healthcare systems record ECG scans of patients with various heart conditions/anomalies. However, due to the need to protect patients' privacy and confidentiality, these data often cannot be shared, meaning it is difficult to produce accurate ECG-based prediction systems for cardiac conditions, especially for conditions that are not prevalent nor commonly reported. However, using Alberta data set, we might be able to produce ECGs with certain abnormalities using a generative algorithm, such as variational autoencoders (VAE), while retaining the privacy of individual patient information [24]. We can then add these synthetic ECGs to the real-world labeled ECG training set (such as CPSC dataset [12]), to produce a model that (potentially) performs better on an external (real) test set.

To diagnose an ECG abnormality, both morphologies of a single beat (R peaks, presence of P wave, ..) and rhythm (combination of multiple beats), it is useful to consider both [7]. In this regard, Jang *et al.* [8] used unsupervised convolutional VAE to encode input ECGs into 60 features through the reconstruction of the rhythm of lead of ECG (both morphologies of single beat and the rhythm) collected from 1278 patients. van de Leur *et al.* [9] learned a VAE (from 1.1 million ECGs) to encode 12-lead ECG signals of a single beat into 21 learned features, then used these learned features for downstream tasks of detection of reduced ejection fraction, and 1-year mortality. However, the focus of both studies was on using the extracted ECG embedding for a downstream task. To the best of our knowledge, no study explored ways to generate synthetic ECGs using VAE and used them as a data augmentation method.

These studies reached good arrhythmia classification performance using VAE-encoded features from ECGs. However, their models can generate either multiple beats (rhythm) of a single lead or a single beat of a 12-lead ECG signal. Both morphologies and rhythms of ECG signals are important for the diagnosis of ECG abnormalities as different abnormalities express their characteristics in different leads or they are related to rhythm rather than the morphology of a single beat [10]. In our study, we use a generative model (VAE) to learn the rhythm of 12-lead ECG signals using a large dataset of 244,077 patients admitted to hospitals in the Alberta dataset (described below) between February 2007 and April 2020, each labeled with the cardiovascular diagnoses identified with specified ICD-10 codes. We then use this trained model to generate new ECGs, each with one or more of these specified diagnoses. We then identified ways to use this generated ECGs to improve the performance of multi-label classification, using publicly available 12-lead ECG dataset with various abnormalities [12].

II. Methods

The ECG data for this study received approval from the Health Research Ethics Board of University of Alberta.

A. Alberta ECG Dataset

We have access to 244,077 patients, where each had 12-lead ECG traces (each collected using 500 Hz frequencies for a duration of 10 seconds) along with 22 common measurements of ECG provided by the Philips IntelliSpace ECG system. Sun *et al.* [4] provide additional details of this dataset. We then applied the Butterworth pass filter (provided by Neurokit library [11]) to remove baseline noise, then normalized the ECG signals over time with the z-score normalization method. This maps each (clean normalized) ECG signal to 4096 real values. We divide the dataset into a 60% development set (train + validation) (964,741 ECG — 146,466 Patients) and a 40% test set (640,527 ECG — 97,631 Patients-disjoin from the training patients).

B. China Physiological Signal Challenge 2018 Dataset

A recent challenge competition provided the CPSC 2018 dataset [12] consisting of 12-lead ECGs collected from 11 hospitals using the frequency of 500 Hz, each with one or more of 9 possible labels: Sinus Rhythm (SR), Atrial Fibrillation (AFIB), First-degree Atrioventricular Block (I-AVB), Left Bundle Branch Block (LBBB), Right Bundle Branch Block (RBBB), Premature Atrial Contraction (PAC), Premature Ventricular Contraction (PVC), ST-segment Depression (STD), and ST-segment Elevated (STE). More detailed description of the CPSC dataset is discussed in Supporting Information (SI) (Section S1). We divided the training CPSC dataset into 80 % training (5503 ECGs), 10 % validation (687 ECGs), and 10% test set (687 ECGs). The test set was fixed for all experiments (with and without data augmentation).

C. Variational Autoencoder Model

We used the VAE architecture and code provided by van de Leur *et al.* [9], but modified it to train and reconstruct the rhythm of 12-lead ECGs of the Alberta Dataset. The architecture is based on β-VAE [13], which is a special case of VAE that adds the hyperparameter of β to the loss function (divergence loss term) to learn disentangled embeddings (see Figure S2). Following van de Leur *et al.* [9], we set β-VAE's adjustable parameters, the number of embeddings and β, to 32 and 8, respectively. β-VAE's consists of three components: encoder, bottleneck and decoder. During the learning process, the 12-lead ECGs were fed into the encoder section of β-VAE, which compresses this signal into 32 means and 32 variances, with the goal of being able to use this generated 32-tuple to reconstruct the input ECG signal (in the decoder). [1]

D. Data Generation

We will learn one set of 64 β-VAE parameters for each cardiovascular diagnosis (32 [mean, variance] pairs) from Alberta ECG Dataset. We will then use this learned model to generate new (realistic) synthetic ECG signals. During data generation (Figure 1), we froze the layers weights of both the encoder and decoder. Then, we feed selected 12-lead ECGs X with a specified abnormality into the encoder. Then, using the means and variance, we draw a sample (Z) from the Gaussian distribution and feed it into the decoder. This generates a new 12-lead ECG (associated with the same abnormality as the one fed into the encoder), which we can then use to augment the CPSC 2018 dataset.

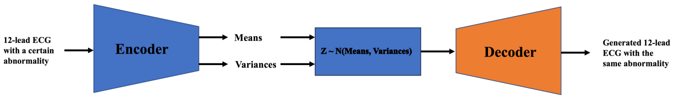

12-lead ECG with a certain abnormality
Encoder
Means
Variances
Z ~ N(Means, Variances)
Decoder
Generated 12-lead ECG with the same abnormality

Fig. 1. Schematic of ECG generation using trained β-VAE. The weights of both the Encoder and Decoder layers were fixed during data generation.

E. Learning Algorithm

We use the InceptionTime model [14] for multi-label classification of ECG abnormalities of CPSC 2018 Dataset, based on the TSAI [15] public library that implements many state-of-the-art algorithms for time series tasks. Using a training data set, the model was trained for 200 epochs with early stopping. The mean F1 score was monitored for early stopping, and if the validation's F1 score was not improved for 50 epochs, the training was stopped, and the model was stored for the inference stage on the test set.

F. Evaluation

Our evaluations are divided into two parts: (1) Evaluation of the quality of β-VAE's learned embeddings using the Alberta (AB) Dataset (discussed in SI, section S2), and (2) Evaluation of different data augmentation methods based on the downstream prediction error of the classifiers learned using that data, for the task of multi-label classification of ECG abnormalities of CPSC 2018. Here, we consider using ECGs generated by a model learned from AB, data addition of AB ECGs, and oversampling of ECGs of CPSC 2018 dataset.

G. Evaluation of Various Data Augmentation Methods

Since β-VAE is a generative model, we can use it to generate ECG instances, which we can use to augment an existing dataset, to see if a classifier learned from this extended dataset can improve the performance compared with the original real dataset. To capture the role of generated ECGs on the performance of the classifier, we design multiple experiments that use various data augmentation techniques; see Table I. First,

[1] The code base for training the β-VAE model used in this study is available at here

we measure the performance of the classifier without any data augmentation method as a baseline experiment (CPSC_NA). Then, we augment the train+validation dataset with various augmentation; ECGs from AB dataset as a positive control for the contribution of augmented ECGs (AB_Orig_STE), VAE synthetic ECGs as a target experiment to evaluate the effect of ECG generated data (ABVAE_Gen_STE), and oversampled CPSC ECGs as a negative control (CPSC_OS_STE). These generated ECG signals were then added with a ratio of 90 % into training and 10 % into the validation set of the CPSC 2018 Dataset. For a fair comparison, we fixed the test set for all experiments. For all experiments, we used the F1 score of the test set to evaluate the model's performance. To determine if the results of the multi-label classification of ECG abnormalities of CPSC 2018 are statistically significant, we used the bootstrapping method: The test set was sampled 10,000 times with random replacement sampling. Then, the 95% confidence intervals of the F1 score for each label were calculated. Additionally, we calculated a bootstrapped difference of means. We also calculated the 95% confidence interval of the difference of means per abnormality and for the means of abnormalities for each label and experiment to check if the observations are statistically significant.

TABLE I
THE EXPERIMENTS PERFORMED IN THIS STUDY.

Experiment Name	Training Sample Size (# of ECGs)	Description
CPSC_NA	6190	No data augmentation as baseline experiment
AB_Orig_STE	6190 + 1072 AB ECGs	1072 real ECGs with STE from AB dataset as a positive control experiment
ABVAE_Gen_STE	6190 + 1072 VAE generated From ABVAE	1072 AB VAE generated ECGs with STE abnormality as a target experiment
CPSC_OS_STE	6190 + 1072 oversampled ECGs from CPSC	1072 oversampled ECG from CPSC dataset as a negative control experiment.

III. Results

In the following, we evaluate the role of synthetic ECGs on the performance of a classifier for the task of multi-label classification using CPSC 2018.

A. Performance of Multi-Label Classification of ECG Abnormalities for CPSC 2018

Figure 2 shows the models' performance on the test set. As a first data augmentation experiment (CPSC_NA), we selected the label with the lowest F1 score (STE), then selected new raw instances from the AB dataset. In particular, we used AB instances whose STEMI label was a negative diagnosis for all other labels (72 cases had this condition), and 1000 samples of STEMI that had a negative diagnosis for at least the labels of the CPSC 2018 ECG dataset (AB_Orig_STE). Note that our dataset labels are limited to only 15, and selected ECGs might have other possible abnormalities not covered by our set of labels.

The results show that the addition of synthetic data (generation or augmentation) increased the F1 score performance of STE compared to the models trained on just the original dataset. The addition of raw AB ECGs of patients with STE (AB_Orig_STE) labels had the highest performance (0.0890[0.0597-0.1185], mean pairwise difference in F1 scores followed by 95% confidence interval of the mean pairwise difference,) for STE diagnosis compared with ABVAE_Gen_STE (0.0463[0.0267-0.0659]) or CPSC_OS_STE (0.0447[0.0135-0.0760]) approaches. However, data addition had a mixed effect on the model's performance of other labels. For the AB_Orig_STE experiment, we observed statistically significant changes in SR, AFIB, 1-AVB, PAC, PVC, STD and STE, an statistically insignificant difference in LBBB and RBBB. AB_Orig_STE data has better performance in AFIB, 1-AVB, PVC, STE and worse in SR, PAC, and STD.

For the AB_Gen_VAE, we observed statistically significant changes in the performance of SR, AFIB, 1-AVB, LBBB, PVC, STD, and STE labels. An insignificant difference in PAC and RBBB, a significant decrease in the performance for the STD label, and a slightly worse performance on AFIB labels. For the CPSC_OS_STE experiment, we observed a significant difference in performance for AFIB, 1-AVB, LBBB,SR, PVC, STD, and STE. Performance was not significant for RBBB and PAC. Oversampling significantly decreased the performance for STD, but did improve performance for AFIB, 1-AVB, LBBB, SR, PVC, and STE. Most performance increases were small but LBBB had significant improvement of 0.03 to 0.05 in F-1. These results suggest that adding either β-VAE, or real ECG signals of the AB dataset, to the training dataset led to models that had an overall better performance improvement compared with oversampling of STE ECG signals from the CPSC 2018 ECG dataset.

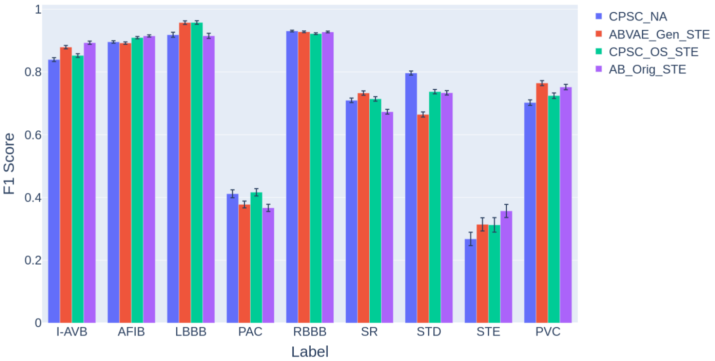


Fig. 2. Model performance of classification of ECG abnormalities for CPSC 2018 dataset. The error bar shows the upper and lower 95% confidence intervals.

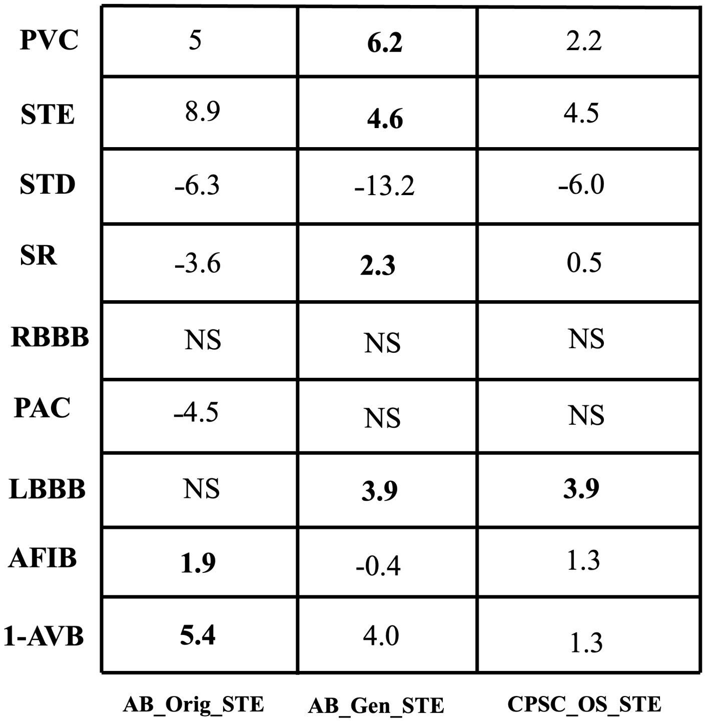
IV. Discussion

We used β-VAE to develop a generative model of the rhythm of 12-lead ECG signals. To the best of our knowledge, this is the first study that was able to learn the rhythm of 12 lead signals using β-VAE. Using β-VAE, other studies were able to learn either the rhythm of 1-lead ECG signals [8] or 1 beat of 12-lead signals [9]. (Note they also used the learned embeddings for downstream tasks). Here, using generated ECG data from learned β-VAE based on a large ECG dataset of AB Hospitals, we investigate the role of synthetic data to help learn models that can classify ECG abnormalities of CPSC 2018 ECG dataset. We focused on the comparison of the model's performance under different numbers of AB β-VAE generated ECGs, over-sampling of ECGs, and addition of new ECGs obtained from the AB Hospitals Dataset (Figure 3). We found that AB β-VAE generated ECGs with STE abnormality not only were able to improve the model's performance on the STE label of the test set but also improve the model's performance on 4 other labels. The performance of oversampling the STE label also improved the model's performance of the STE label, but its positive effect on the performance of other labels was less than β-VAE generated data. For the STE label, among AB dataset β-VAE generated ECG data and the addition of AB original ECG data, the AB original ECGs improved the model's performance of the STE label by $\sim$9 %, while the AB β-VAE generated ECGs improved it by $\sim$5 %. We assume this lower performance (of AB β-VAE generated ECGs compared with the AB original ECG data) is because the reconstructed ECGs were not perfect and there was some information loss. However, this reduction might be acceptable, as it means the ECGs used do not compromise the patients' privacy.

The beneficial effect of ECG data generation on the model's performance was previously introduced by other studies, which used generative adversarial networks (GAN) to generate synthetic ECG data. Wang *et al.* [17] used a modified version of GAN called auxiliary classifier generative adversarial network (ACGAN) to generate synthetic data. Their method requires first identifying the R peaks of the signal and concatenation of 5 generated heartbeats as a sample (12 * 1500 [lead, datapoints]). They used the CPSC 2018 dataset, where they segmented the original ECG data into lower lengths that resulted in 13754 samples rather than original 6877 samples. Then, they selected 50 instances from each label as a test set. These generated ECG data improved the performance of the classifier in the test set for all labels, compared with models that were trained with no data-generated ECG. Since they segmented the original dataset into short lengths, we cannot directly compare our classifier performance with this study. Others also used GAN-based methods to generate synthetic ECG data and observe an improvement of data generation over their baseline model using other ECG datasets [17] [18].

There are some limitations associated with our study. While our β-VAE model was able to learn the 12-lead signals, the predictive ability of these learned embedding was lower than 22 ECG Global Measurements. The focus needs to be shifted to finding algorithms that are able to better encode ECG domains, and it will eventually enhance the performance of the classifier (using embeddings as features) for the downstream task.

Also, we assessed the similarity between real and synthetic ECG signals using a common metric, such as Mean Squared Error (MSE). However, it may not adequately capture the nuanced ECG waveform characteristics relied upon by expert ECG readers for traditional ECG interpretation. Future work will involve collaboration with ECG experts who will conduct a qualitative evaluation of the generated ECGs. This collaboration aims to ensure that our synthetic ECG data closely resembles real ECGs, particularly in relation to their original labels or abnormalities. Furthermore, we intend to include quantitative metrics (e.g., P-wave duration) to compare synthetic ECGs with real ones, further validating the quality of the generated ECGs.



	AB_Orig_STE	AB_Gen_STE	CPSC_OS_STE
PVC	5	**6.2**	2.2
STE	8.9	**4.6**	4.5
STD	-6.3	-13.2	-6.0
SR	-3.6	**2.3**	0.5
RBBB	NS	NS	NS
PAC	-4.5	NS	NS
LBBB	NS	**3.9**	**3.9**
AFIB	**1.9**	-0.4	1.3
1-AVB	**5.4**	4.0	1.3

Fig. 3. Mean F1 score (%) differences between data augmentation approach and original real CPSC for each label. (NS represents "not statistically significant".)

V. Conclusion

Here, we have trained an unsupervised β-VAE to generate 12-lead ECG signals based on a large dataset of AB ECGs. This framework can generate synthetic ECGs with a certain abnormality. Then, we evaluated the quality of AB β-VAE generated ECG data by seeing whether adding them as additional training data of ECG abnormalities, to the publicly available CPSC 2018 ECG dataset. We found that adding generated AB ECG data not only improves the performance of the targeted label (STE label) but also significantly improves the performance of 4 other labels, while having no effect or slightly negative effect on the performance of other labels. Future studies are required to evaluate the beneficial effect of AB β-VAE generated ECG data on other datasets. Also, more effort is required to develop stronger β-VAE frameworks that can encode the ECGs into richer embeddings and eventually help to build stronger classifiers.

Acknowledgment

This work was supported by NSERC Discovery, IVADO, CIFAR through CCAI Chair funding and by the Alberta Machine Intelligence Institute (Amii)

References

[1] Dahlöf, B., “Cardiovascular disease risk factors: epidemiology and risk assessment,” Am. J. Cardiol., vol. 105, pp. 3A–9A, 2010

[2] Ribeiro, A.H., Ribeiro, M.H., Paixão, G.M., Oliveira, D.M., Gomes, P.R., Canazart, J.A., Ferreira, M.P., Andersson, C.R., Macfarlane, P.W., Meira Jr, W. and Schön, T.B., “Automatic diagnosis of the 12-lead ECG using a deep neural network,” Nat. Commun., vol 11, pp. 1760, 2020

[3] Yıldırım, Ö., Pławiak, P., Tan, R.S. and Acharya, U.R, “Arrhythmia detection using deep convolutional neural network with long duration ECG signals,” Comput. Biol. Med., vol. 102, pp. 411–420, 2018

[4] Sun, W., Kalmady, S.V., Sepehrvand, N., Salimi, A., Nademi, Y., Bainey, K., Ezekowitz, J.A., Greiner, R., Hindle, A., McAlister, F.A. and Sandhu, R.K., “Towards artificial intelligence-based learning health system for population-level mortality prediction using electrocardiograms,” NPJ Digit. Med., vol. 6, pp. 21, 2023

[5] Winter, E., “The shapley value,” Handbook of game theory with economic applications, vol. 3, pp. 2025–2054, 2002

[6] Ribeiro, M.T., Singh, S. and Guestrin, C., “Why should i trust you?” Explaining the predictions of any classifier,” Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144, 2016

[7] Berkaya, S.K., Uysal, A.K., Gunal, E.S., Ergin, S., Gunal, S. and Gulmezoglu, M.B., “A survey on ECG analysis,”, Biomed. Signal Process. Control, vol. 43, pp. 216–235, 2018

[8] Jang, J.H., Kim, T.Y., Lim, H.S. and Yoon, D., “Unsupervised feature learning for electrocardiogram data using the convolutional variational autoencoder,” PLoS One, vol. 16. pp. e0260612, 2021

[9] Van de Leur, R.R., Bos, M.N., Taha, K., Sammani, A., Yeung, M.W., van Duijvenboden, S., Lambiase, P.D., Hassink, R.J., van der Harst, P., Doevendans, P.A. and Gupta, D.K., “Improving explainability of deep neural network-based electrocardiogram interpretation using variational auto-encoders,” EHJDH, vol. 3, pp. 390–404, 2022

[10] Liu, X., Wang, H., Li, Z. and Qin, L., “A survey on ECG analysis,” KBS, vol. 227, pp. 107187, 2021

[11] Makowski, D., Pham, T., Lau, Z.J., Brammer, J.C., Lespinasse, F., Pham, H., Schölzel, C. and Chen, S.A., “NeuroKit2: A Python toolbox for neurophysiological signal processing.,” Behav. Res. Methods, vol. 53, pp. 1689–1696, 2021

[12] Liu, F., Liu, C., Zhao, L., Zhang, X., Wu, X., Xu, X., Liu, Y., Ma, C., Wei, S., He, Z. and Li, J., “An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection,”, J. Med. Imaging & Health Infor., vol. 8, pp. 1368–1373, 2018.

[13] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S. and Lerchner, A.,“beta-vae: Learning basic visual concepts with a constrained variational framework,”, In International conference on learning representations., 2016

[14] Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D.F., Weber, J., Webb, G.I., Idoumghar, L., Muller, P.A. and Petitjean, F., “Inceptiontime: Finding alexnet for time series classification,”, Data Min Knowl Discov, vol 34, pp. 1936–1962, 2020

[15] I. Oguiza, “tsai - A state-of-the-art deep learning library for time series and sequential data,”, 2022

[16] Wang, P., Hou, B., Shao, S. and Yan, R., “ECG arrhythmias detection using auxiliary classifier generative adversarial network and residual network,”IEEE Access, vol 7, pp. 100910–100922, 2019

[17] Ma, S., Cui, J., Chen, C.L., Chen, X. and Ma, Y., “An effective data enhancement method for classification of ECG arrhythmia,”Measurement, vol 203, pp. 111978, 2022

[18] Zhang, Y.H. and Babaeizadeh, S.,“Synthesis of standard 12-lead electrocardiograms using two-dimensional generative adversarial networks,”, CJC, vol 69, pp. 6–14, 2021

[19] Jagannathan, R., Patel, S.A., Ali, M.K. and Narayan, K.V., “Global updates on cardiovascular disease mortality trends and attribution of traditional risk factors,”Curr. Diab. Rep., vol 19, pp. 1-12, 2019.

[20] Hong, S., Zhou, Y., Shang, J., Xiao, C. and Sun, J., “Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review.,” 2020. Comput. Biol. Med., vol 122, p.103801.

[21] Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A. and Bennett, K.P.,“Generation and evaluation of privacy preserving synthetic health data. ,” Neurocomputing, vol. 416, pp.244-255, 2020.

[22] Jagannathan, R., Patel, S.A., Ali, M.K. and Narayan, K.V., “Global updates on cardiovascular disease mortality trends and attribution of traditional risk factors.,”, Current diabetes reports, vol 19, pp.1-12. 2019.

[23] Hong, S., Zhou, Y., Shang, J., Xiao, C. and Sun, J., “Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review.,”, Computers in biology and medicine, vol 122, p.103801, 2020.

[24] Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A. and Bennett, K.P.,“Generation and evaluation of privacy preserving synthetic health data.,”, Neurocomputing, vol 416, pp.244-255, 2020.

Supporting Information

VI. Section S1

As discussed in the Method section of the main manuscript, CPSC 2018 dataset consists of 9 possible labels: Sinus Rhythm (SR), Atrial Fibrillation (AFIB), First-degree Atrioventricular Block (I-AVB), Left Bundle Branch Block (LBBB), Right Bundle Branch Block (RBBB), Premature Atrial Contraction (PAC), Premature Ventricular Contraction (PVC), ST-segment Depression (STD), and ST-segment Elevated (STE). The dataset was previously divided into the training set (6877 instances (female: 3178; male: 3699)) and test set (2954 instances (female: 1416; male: 1538)) by the competition, where the test set, which is still not public, was used to rank participants. Table S1 shows the number of training set recordings for each label. The majority of the ECG training data has only 1 label (6400 samples), while 477 samples have multiple abnormalities in their ECGs.

TABLE S1

Total number of ECGs in the training set for various labels.

Challenge Set	**Label**	**Total Number of Recordings**
Training	SR	918
	AFIB	1098
	1-AVB	704
	LBBB	207
	RBBB	1695
	PAC	574
	PVC	653
	STD	826
	STE	202
Total		**6877**

VII. Section S2: Evaluation of β-VAE Embeddings

To evaluate the quality of learned embeddings, we use the embeddings for the task of multi-label classification of

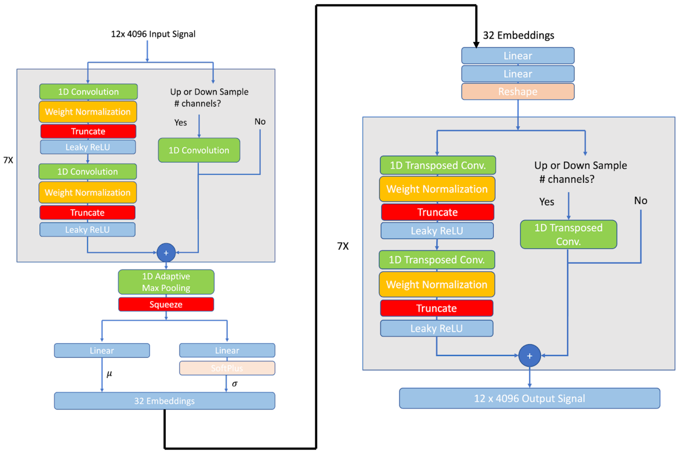


Fig. S1. The architecture of β-VAE used in this study.

cardiovascular diagnosis of Alberta (AB) ECG dataset. To generate the 32 ECG embeddings, we feed the 12-lead ECGs of AB ECG dataset into the encoder section of trained β-VAE (Figure S1), which produces 32 pairs [means, variances], which we use the means to represent that signal. Afterwards, we ran the gradient boosted tree ensembles (XGBoost) model on these instances (from the train + validation set), along with age and sex, to learn 15 models – for one versus all classifications of each label. We then evaluated the performance of each model using the area under the receiver operating curve (AUROC) and the F1 score of the test set.

A. Performance of Multi-Label Classification of cardiovascular diagnoses of AB ECG Dataset

The VAE reconstructed the AB 12-lead ECG signals with various error levels for different training ECG signals. We also calculate the correlation of these 32 embeddings with 22 ECG measurements provided by a Philips machine (Figure S2). These correlations can provide an explanation for the characteristics of learned embeddings and their relation with well-defined ECG measurements. Due to the unsupervised nature of β-VAE, the model might learn the characteristics of some ECG labels more than other labels depending on the number of training instances with that label. To evaluate the quality of learned ECG features for different labels, we used these features, as well as age and sex, for the task of multi-label classification of cardiovascular diagnoses. To this end, we used XGBoost to create 15 independent models, one for each type of diagnosis (Table S2).

If we select AUROC = 0.70 as a threshold for reasonable learning performance, 9 labels – ST Elevation Myocardial Infarction (STEMI), Heart Failure, Unstable Angina, Atrial Fibrillation, Ventricular Tachycardia, Atrioventricular Block, Pulmonary Hypertension, Hypertrophic Cardiomyopathy, and Hypertrophic Cardiomyopathy – have the performance above this threshold, suggesting that based on this learned β-VAE, these labels might be more suitable candidates for data generation as compared to other 6 labels.

TABLE S2
THE MULTI-LABEL CLASSIFICATION OF 15 CARDIOVASCULAR DIAGNOSES USING THE ALBERTA ECG DATASET USING 32 TCN BASED β-VAE EMBEDDINGS, 22 ECG GLOBAL MEASURMENTS(GMS) FROM PHILIPS MACHINE. ALL FEATURES/EMBEDDINGS INCLUDE AGE AND SEX AS ADDITIONAL FEATURES.

Label	32 Embeddings		22 ECG GMs	
	AUROC	F1	AUROC	F1
NSTEMI	0.65	0.22	0.77	0.33
STEMI	0.74	0.25	0.88	0.49
Heart Failure	0.77	0.33	0.83	0.35
Unstable Angina	0.73	0.14	0.76	0.18
Atrial Fibrillation	0.75	0.31	0.72	0.18
Ventricular Tachycardia	0.72	0.09	0.77	0.12
Cardiac Arrest	0.64	0.06	0.74	0.09
Supraventricular Tachycardia	0.62	0.10	0.60	0.08
Atrioventricular Block	0.84	0.23	0.89	0.31
Pulmonary Embolism	0.61	0.04	0.69	0.11
Aortic Stenosis	0.68	0.05	0.80	0.09
Pulmonary Hypertension	0.70	0.05	0.77	0.11
Hypertrophic Cardiomyopathy	0.70	0.03	0.86	0.11
Mitral Valve Prolapse	0.62	0.02	0.72	0.04
Mitral Valve Stenosis	0.61	0.01	0.76	0.02

VIII. SECTION S3: THE PAIRWISE DIFFERENCES BETWEEN AUGMENTED MODELS AND THE BASELINE MODEL.

Tables S3 through S7 display the pairwise differences in mean F1 scores for each label and pair of models. We also computed the 95% confidence interval for the difference in means for each abnormality, as well as the average difference across all abnormalities for each label. If the lower and upper bounds of the confidence interval have different signs, indicating they crossed zero, it shows that the observation is not statistically significant.

TABLE S3

The pairwise differences between AB_Orig_STE and CPSC_NA models.

Label	mean_F1	CI_upper	CI_lower
SR	-0.0363	-0.0277	-0.0469
AFIB	0.0191	0.0243	0.0138
1-AVB	0.0537	0.0618	0.0457
LBBB	-0.0033	0.008	-0.014
RBBB	-0.0030	0.0008	-0.0068
PAC	-0.0448	-0.0277	-0.0620
PVC	0.0497	0.0619	0.0376
STD	-0.0631	-0.0536	-0.0726
STE	0.0890	0.1185	0.0597

TABLE S4

The pairwise differences between ABVAE_Gen_STE and CPSC_NA models.

Label	mean_F1	CI_upper	CI_lower
SR	0.0233	0.0296	0.0169
AFIB	-0.0037	-0.0006	-0.0069
1-AVB	0.0398	0.0449	0.0348
LBBB	0.0388	0.0442	0.0334
RBBB	-0.0025	-0.00001	-0.0005
PAC	-0.0340	-0.0296	0.0169
PVC	0.0623	0.0698	0.0548
STD	-0.1322	-0.1246	-0.1398
STE	0.0463	0.0659	0.0267

TABLE S5

The pairwise differences between CPSC_OS and CPSC_NA models.

Label	mean_F1	CI_upper	CI_lower
SR	0.0046	0.0145	-0.0053
AFIB	0.0135	0.0188	0.0081
1-AVB	0.0130	0.0216	0.0044
LBBB	0.0390	0.0488	0.0293
RBBB	-0.0085	-0.0047	-0.0122
PAC	0.0049	0.0222	-0.0123
PVC	0.0221	0.0344	0.0098
STD	-0.0593	-0.0497	-0.0689
STE	0.0447	0.0760	0.0135

TABLE S6

The pairwise differences between ABVAE_Gen_STE and CPSC_OS models.

Label	mean_F1	CI_upper	CI_lower
SR	0.0187	0.0286	0.0088
AFIB	-0.0172	-0.0118	-0.0226
1-AVB	0.0268	0.0349	0.0188
LBBB	-0.0003	0.0079	-0.0084
RBBB	0.0060	0.0098	0.0022
PAC	-0.0390	-0.0231	-0.0549
PVC	0.0402	0.0519	0.0285
STD	-0.0729	-0.0621	-0.0836
STE	0.0016	0.0328	-0.0296

TABLE S7

The pairwise differences between AB_Orig_STE and CPSC_OS models.

Label	mean_F1	CI_upper	CI_lower
SR	-0.0409	-0.0306	-0.0513
AFIB	0.0056	0.0106	0.0005
1-AVB	0.0407	0.0487	0.0328
LBBB	-0.0423	-0.0324	-0.0522
RBBB	0.0055	0.0093	0.0016
PAC	-0.0498	-0.0335	-0.0661
PVC	0.0276	0.0397	0.0156
STD	-0.0038	0.0062	-0.0138
STE	0.0444	0.0752	0.0135