Multimodal learning and fusion from 12-Lead ECG

Anita Salvador[1], Nicolas Duchateau[2,3], Badih Ghattas[1]

[1] Aix-Marseille University, Marseille, France
[2] INSA-Lyon, Univ. Lyon 1, CNRS, Inserm, CREATIS, UMR 5220, U1294, Villeurbanne, France
[3] Institut Universitaire de France (IUF)

Abstract

Learning meaningful low-dimensional representations from medical data is essential for disease characterization and interpretability. For ECG signals, where each lead provides both redundant and complementary information. We aim at quantifying the value of explicitly considering this structure with multimodal variational autoencoders (MVAEs), compared to simpler early or late fusion with classical VAE.

Each ECG lead was considered a separate modality. We used a MVAE tailored for ECG information fusion to learn a shared latent space capturing common and complementary information across all leads, while encoding/decoding each lead separately. We evaluated the quality of the learnt representations through their organization in the latent space and the separability between controls and known diseases. Comparisons included early fusion and late fusion.

Modeling 12-lead ECGs as a multimodal signal within a MVAE framework enables the learning of richer representations that consider redundancy and complementarity across leads, of high relevance for the characterization of subtle alterations of ECG shape due to disease.

1. Introduction

Electrocardiograms (ECGs) provide complex and multidimensional information about cardiac activity, with morphological alterations associated with different cardiac conditions. In twelve-lead ECGs, this complexity is further increased by the multiple synchronized views of the same cardiac activity, carrying both redundant and complementary information. Learning a compact low-dimensional representation of these signals can facilitate the characterization of patient populations by summarizing this complex information into a simpler representation in which different pathological patterns can be identified and separated.

Variational Autoencoders (VAEs) [1] provide an unsupervised framework for learning such representations directly from high-dimensional data, without requiring diagnostic labels during training. Their formulation can also be extended to multimodal data [2], making them particularly relevant for twelve-lead ECG analysis. Different strategies can be used to handle the multiple ECG leads: processing them jointly within a single VAE (early fusion), learning each lead independently before combining their representations (late fusion), or learning a shared representation while explicitly preserving their multimodal structure with a Multimodal VAE (intermediate fusion).

In this work, we compare these three strategies for learning low-dimensional representations of twelve-lead ECGs. We evaluate whether explicitly accounting for their multimodal structure improves the characterization of ECG populations, through the representation of clinically relevant information and the separability of healthy controls and patients with different conduction disorders.

2. Methods

2.1. Data description

The study was conducted on twelve-lead electrocardiograms extracted from the public 2020 PhysioNet Challenge database [3]. Each recording consists of twelve synchronized leads (I, II, III, aVR, aVL, aVF, V1–V6), providing different views of cardiac electrical activity, which we consider different inputs for a multimodal analysis. Patients were divided into three subgroups according to their diagnosis: healthy controls (N=917), left bundle branch block (LBBB, N=176), right bundle branch block (RBBB, N=1646). These pathologies were selected because they induce characteristic morphological alterations that affect the twelve leads differently, making them particularly relevant for evaluating multimodal representation learning approaches.

2.2. Preprocessing

For each patient, cardiac cycles were detected and temporally aligned according to the QRS [4]. The extracted

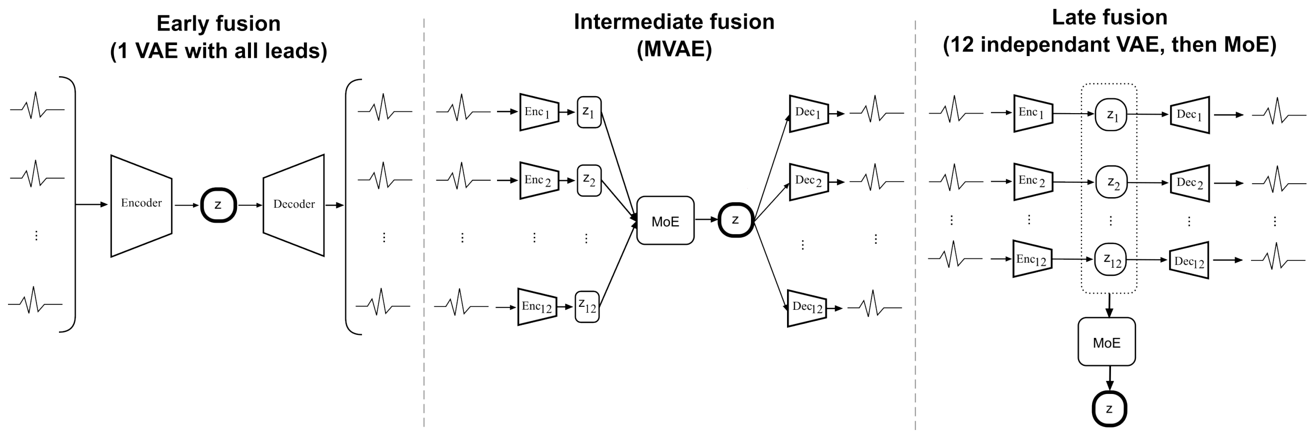

Figure 1. Shematic representation of th three fusion strategies considered in this work. Early fusion, where the twelve ECG-leads are treated as twelves input channels of a single VAE. Intermediate fusion (MVAE) considered each ECG-lead as one channel each one having its specific encoder and decoder, but sharing a common latent space. Late fusion learns independantly from each lead a specific latent space that are combine into a common one after the trainning phase. In this figure, z denotes the shared latent representation, while z_i denote the lead-specific latent representation associated with the i^{th} ECG-lead.

beats were resampled to a common temporal resolution and a representative heartbeat was obtained by taking the average signal across cycles. Since ECG amplitudes may vary substantially across patients and derivations, in each lead the amplitude was normalized independently before representation learning. In our framework, the twelve ECG leads were explicitly modeled as distinct modalities carrying redundant and also complementary information about cardiac activity.

2.3. Learning the representation

Three representation-learning strategies were compared to study the impact of multimodal modeling. In the early-fusion approach, the twelve leads were concatenated and processed jointly as twelve input channels of a single one-dimensional convolutional variational autoencoder (VAE). In the intermediate-fusion approach, we designed a multimodal variational autoencoder (MVAE) based on the Mixture-of-Expert formulation [2] in which each lead was encoded and decoded by a dedicated one-dimensional convolutional branch, while the latent distributions were combined through a mixture-of-experts mechanism to learn a shared latent space capturing information from all the leads. Finally, in the late-fusion approach, twelve independent single-lead VAEs were trained separately and their latent representations were fused afterward, also through a mixture-of-experts method, to produce a common embedding. All three methods are represented in Figure 1.

All architectures relied on one-dimensional convolutional layers to account for the temporal structure of ECG signals and were trained in an unsupervised manner, that is during the trainning the model had only access to the ECG leads, not the diagnostic labels. The models were implemented based on the multiviewae librairy from Aguila [5]. The latent space dimensionality was set to 16, corresponding to the minimum number of principal components required to explain at least 95% of the variance in a PCA. The dataset was split into training, validation, and testing representing respectivly 70%, 20% and 10% of the patients. Models were trained for 500 epochs, with a batch size of 64 with the Adam optimizer.

2.4. Evaluation methods

The learned latent spaces were evaluated according to their ability to organize and separate groups characterized by different ECG morphologies, including control group, LBBB, RBBB. Representation quality was assessed through both qualitative and quantitative analyses. To measure the latent organization we use an index taking into acount the position in the latent space and an external feature such as the QRS-length. This index is calculated after operating an O-PLS on the latent space with respect to the QRS-lenght to obtain the principal direction of evolution of this external feature, then we calculate the correlation coefficient between the coordinates in this direction wich become our index of the latent space organisation of this specific feature. and the feature Class separability was quantified using normalized separability ratios and pairwise Hotelling T^2 tests. Together, these metrics allowed us to compare the ability of each fusion strategy to capture discriminative and physiologically meaningful information from the twelve-lead ECG alone.

3. Results

Table 1. Correlation indexes between the latent representation and the QRS duration for the different fusion strategies. Higher values indicate stronger correlations.

	VAE early fus.	MVAE inter. fus.	12 VAE late fus.
All	**0.27**	0.24	0.18
Control	**0.30**	**0.30**	0.18
Patho	**0.41**	0.28	0.30
RBBB	**0.40**	0.18	0.25
LBBB	***0.71***	*0.65*	*0.62*

Table 2. Hotelling T^2 statistics for the different fusion strategies. Higher values indicate better class separability in the latent space.

	VAE early fus.	MVAE inter. fus.	12 VAE late fus.
Control vs Patho	144	**179**	101
Control vs LBBB	372	**435**	368
Control vs RBBB	**390**	338	382
LBBB vs RBBB	315	**439**	245

Table 1 reports the correlation between the learned representations and QRS duration, a clinically relevant descriptor, specially in the RBBB case. Overall, the early-fusion approach tends to exhibit higher correlations, while the MVAE achieves slightly lower values. The particularly high correlations observed for LBBB may partly result from its much smaller sample size (ten times smaller than the RBBB group) making these estimates less stable. Although the correlation values differ across methods, they consistently remain moderate, indicating that QRS duration is partially encoded in the learned latent representations. The similar correlation levels obtained with the three fusion strategies further suggest that they all capture this physiological attribute to a comparable extent. These observations indicate that a strong correlation with a single physiological marker is not necessarily associated with a more informative latent representation.

Table 2 provides a complementary evaluation of the learned representations through their ability to separate the different diagnostic groups. While all three approaches successfully discriminate the considered ECG populations, the MVAE achieves the highest Hotelling T^2 values for most comparisons. This suggests that explicitly modeling each ECG lead as an individual modality may help the latent space to capture complementary information beyond QRS duration alone, witch information is present in each ECG-leads, to improved class separability by capturing complementary information distributed across the twelve leads.

4. Conclusion

This work compared three fusion strategies for learning low-dimensional representations from 12(lead ECGs. Overall, all approaches were able to encode relevant information and separate populations characterized by different pathologies, when the MVAE generally provided better class separability.

These results also highlight that the choice of fusion strategy may depend on how information is distributed across modalities. Intermediate fusion explicitly preserves lead-specific processing while learning a shared representation, wich may facilitate the integration of complementary information distributed across ECG leads. However, this architecture may be less advantageous when the relevant information is highly redundant across modalities, as separately encoding similar information in each branch may provide limited aditional benefit compared with processing the leads jointly. Conversly, late fusion is the only strategy considered here where each model is trained without access to the other leads. Its generally lower separability may result from the inalibity to exploit cross-lead relationships during the learning phase.

The distinction between early and intermediaite fusion is more subtle. Rather than identifying one strategy as systematically preferable, their relative benefit may depend on the degree of redundancy and complementarity between the considered modalities. this motivates the exploration of hybrid architectures operating at multiple fusion levels, where strongly redundant leads could first be grouped and jointly encoded, while complementary groups of leands could be combbines throught multimodal fusion. More generally, this suggests that the definition of modalities itself could be adapted to their degree of redundancy and complementarity, rather than systematically treating each ECG lead as a separate modality. Such hierarchical fusion could provide a compromise between preserving modality-specific information and exploiting interaction between closely related signals.

Acknowledgments

This work was partially supported by the PEPR Santé Numérique (ChroniCardio project, 22-PESN-0015).

References

[1] Kingma DP, Welling M. Auto-Encoding Variational Bayes. arXiv December 2013;.

[2] Shi Y, Siddharth N, Paige B, Torr PHS. Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models 2019;.

[3] Perez Alday EA, Gu A, Shah A, Liu C, Sharma A, Seyedi S, Bahrami Rad A, Reyna M, Clifford G. Classification of 12-lead ECGs: The PhysioNet/Computing in Cardiology Challenge 2020. PhysioNet July 2022;URL https://doi.org/10.13026/dvyd-kd57. Version 1.0.2.

[4] Aalizade N. Ecg p qrs t wave detecting matlab code. MATLAB Central File Exchange, 2018. URL https://www.mathworks.com/matlabcentral/fileexchange/66098-ecg-p-qrs-t-wave-detecting-matlab-code. Version 1.0.0.0, Retrieved January 28, 2026.

[5] Aguila AL, Chapman J, Altmann A. Multi-modal variational autoencoders for normative modelling across multiple imaging modalities. URL http://arxiv.org/abs/2303.12706.

Address for correspondence:

My Name
My Full postal address
My E-mail address