On Merging Feature Engineering and Deep Learning for Diagnosis, Risk Prediction and Age Estimation Based on the 12-Lead ECG

Eran Zvuloni, Jesse Read, Antônio H. Ribeiro, Antonio Luiz P. Ribeiro, and Joachim A. Behar

Abstract*—*Objective: **Over the past few years, deep learning (DL) has been used extensively in research for 12-lead electrocardiogram (ECG) analysis. However, it is unclear whether the explicit or implicit claims made on DL superiority to the more classical feature engineering (FE) approaches, based on domain knowledge, hold. In addition, it remains unclear whether combining DL with FE may improve performance over a single modality.** ***Methods:*** **To address these research gaps and in-line with recent major experiments, we revisited three tasks: cardiac arrhythmia diagnosis (multiclass-multilabel classification), atrial fibrillation risk prediction (binary classification), and age estimation (regression). We used an overall dataset of 2.3M 12-lead ECG recordings to train the following models for each task: i) a random forest taking FE as input; ii) an end-to-end DL model; and iii) a merged model of FE+DL.** ***Results:*** **FE yielded comparable results to DL while necessitating significantly less data for the two classification tasks. DL outperformed FE for the regression task. For all tasks, merging FE with DL did not improve performance over DL alone. These findings were confirmed on the additional PTB-XL dataset.** ***Conclusion:*** **We found that for traditional 12-lead ECG based diagnosis tasks, DL did not yield a meaningful improvement over FE, while it improved significantly the nontraditional regression task. We also found that combining FE with DL did not improve over DL alone, which suggests that the FE was redundant with the features learned by DL.** ***Significance:*** **Our findings provides important recommendations on 12-lead ECG based machine learning strategy and data regime to choose for a given task. When looking at maximizing performance as the end goal, if the task is nontraditional and a large dataset is available then DL is preferable. If the task is a classical one and/or a small dataset is available then a FE approach may be the better choice.**

***Index Terms*—12-lead ECG analysis, big data, deep learning, feature engineering, physiological time series.**

I. Introduction

RELIABLE systems capable of assisting clinical decision-making processes may significantly improve diagnosis and consequently reduce healthcare costs. Correspondingly, cardiovascular disease management using machine learning-based 12-lead electrocardiogram (ECG) analysis has been studied extensively over the last two decades [1], [2], [3].

Deep learning (DL) is broadly accepted as an effective and suitable data-driven approach in performing feature extraction for classification or regression. In the computer vision field, it is the dominant or even exclusive approach for many data-driven tasks [4]. This is thanks to the inherent properties of convolutional layers which make them very good extractors of features in natural images [5]. Nevertheless, in 1D physiological time-series, and thus for 12-lead ECG analysis, DL superiority to classical feature engineering (FE) based machine learning approaches is still an open question. It also remains to be elucidated whether combining FE with DL yields better performance and whether FE provides complementary or redundant information to DL. Moreover, the performance of FE versus DL might be task-specific and data regime dependant. Accordingly, understanding what tasks are better suited to a FE or DL approach is important in ensuring the best performance.

Besides model performance, the differences between FE and DL raise several computational and dataset elaboration considerations. An obvious advantage of DL is that it bypasses the need for FE, thus resulting in models that are more computationally efficient at inference. However, large datasets are not systematically available for the development of the models, or their technical development may be very expensive. Therefore, there is a need to investigate whether there is value in developing large datasets for the purpose of developing high-performing DL models, i.e., DL models that are superior to those that have been historically developed and based on FE. Other practical

Manuscript received 15 December 2022; accepted 18 January 2023. Date of publication 25 January 2023; date of current version 20 June 2023. This work was supported in part by the Technion EVPR Fund: Hittman Family Fund under Grant ERANET-3-16881, in part by the Ministry of Health, Israel under Grant 3-17550, in part by the Ministry of Science and Technology, Israel & Ministry of Europe and Foreign Affairs (MEAE) of Israel & the Ministry of Higher Education, Research and Innovation (MESRI) of France, in part by Israel PBC-VATAT, in part by the Technion Center for Machine Learning and Intelligent Systems, and in part by A cloud computing grant from the Israel Council of Higher Education, administered by the Israel Data Science Initiative (IDSI). *(Corresponding author: Joachim A. Behar.)*

Eran Zvuloni is with the Faculty of Biomedical Engineering, Technion-IIT, Israel.

Jesse Read is with the Computer Science Laboratory (LIX), École Polytechnique, Institut Polytechnique de Paris, France.

Antônio H. Ribeiro is with the Department of Information Technology, Uppsala University, Sweden.

Antonio Luiz P. Ribeiro is with the Department of Internal Medicine, Faculdade de Medicina, Telehealth Center, Hospital das Clínicas, Universidade Federal de Minas Gerais, Brazil.

Joachim A. Behar is with the Faculty of Biomedical Engineering, Technion-IIT, Haifa 3200003, Israel (e-mail: jbehar@technion.ac.il).

This article has supplementary downloadable material available at https://doi.org/10.1109/TBME.2023.3239527, provided by the authors.

Digital Object Identifier 10.1109/TBME.2023.3239527

considerations relate to explainability and uncertainty analysis (out-of-distribution detection), or robustness of the DL model, as the DL features might be prone to domain shift and become data-biased.

An attempt to compare how ECG features are extracted through DL versus FE was made by Attia et al. [6]. The authors trained a model to provide sex and age estimations by analyzing 12-lead ECG signals. A neural network modeled the relationship between DL and FE features. Moreover, linear combinations of the DL features were used to represent the engineered features. Both experiments provided partial explainability of the DL features. Yet, the features did not achieve a perfect match by being fully modeled or represented, which implied that some information gaps remained. Another work by Beer et al. [7] focused on the task of binary classification of atrial fibrillation (AF) from a single-lead ECG input. The authors added the Hilbert-Schmidt independence criterion to their loss function, constraining their DL model to learn different features from those provided by FE. This experiment succeeded in generating novel DL features. Nevertheless, the features were not of value for the AF classification task, obtaining poor scores when the classification was confined to them. Interestingly, this work demonstrated how DL can discover similar features to those engineered by humans and that may have taken years to formalize.

Despite the progress in finding associations between FE and DL features for ECG analysis, there is a lack of quantitative analysis with respect to the level of redundancy that exists between these two approaches in performing classification or regression learning tasks. Moreover, in the research of Attia et al. and Beer et al, the dataset sizes used were relatively small, ranging from 12,186 to 100,000 ECG recordings, respectively. Thus, the added value of combing FE and DL in a big data regime remains unclear.

Herein, we sought to benchmark and evaluate the superiority of DL algorithms to classical FE-based machine learning approaches, and also investigate the value of merging FE with DL for 12-lead ECG analysis and in a big data regime. We decided to focus our experiments on 12-lead ECG analysis because it is the most commonly used cardiac examination in clinical practice. Three tasks were considered: 1) multiclass-multilabel classification of 6 cardiac arrhythmia diagnosis, 2) binary classification for AF risk prediction, and 3) regression for age estimation. The different experiments for each task (FE exclusive, DL exclusive, and merged FE+DL) utilized the same pipeline and FE-DL merging approach, while model optimization was task-specific.

II. Methods

In the following section, we describe the: A) datasets used, B) FE types and their setup, C) model pipeline and architecture, D) performance measures used together with the statistical analysis and E) learning curve elaboration.

A. Datasets

1) Telehealth Network of Minas Gerais (TNMG) Dataset: The recordings were part of the TNMG dataset [8], [9], [10]. All the recordings were 7–10 seconds long and were resampled to a sampling frequency of 400 Hz. The ECG recordings were originally used by Ribeiro et al. [8]. Table I shows the exact dataset division with respect to the learning tasks and for the different classes considered in our experiments. For the arrhythmia diagnosis task, all the recordings with arrhythmia labels were used, and the test set consisted of the additional 827 recordings with labels manually annotated by three different electrocardiography-expert cardiologists, as used in the previous research [8]. The six classes of arrhythmias considered were: first-degree atrioventricular block (1dAVb), right and left bundle branch block (RBBB, LBBB), sinus bradycardia (SB), AF and sinus tachycardia (ST). Recordings with none of these labels were treated as normal sinus rhythm (NSR). For the AF risk prediction task, a subset of the dataset was used. The complete experimental setting was described in Biton et al. [11]. Briefly, recordings classified as "Future AF" were those from patients with a baseline recording without an AF label and with a subsequent recording within 5 years with a positive AF diagnosis. These correspond to class C_2 in [11]."Non-AF" patients (class C_1) were those with a baseline recording with no AF documented and no AF documented in any subsequent recording within 5 years [11]. For the age estimation task, the test set followed the work of Lima et al. [12]; thus, single recordings per patients of ages 16-85 were taken from the CODE-15% [10] dataset (a subset of the TNMG dataset). Other recordings with patient age available were used for the train and validation sets.

TABLE I
TNMG DATASET AND DATA SPLITTING FOR THE THREE TASKS

Task	Class	Train	Validation	Test	Total
Arrhythmia diagnosis	NSR	2,029,146	41,412	681	2,071,239
	1dAVb	25,555	522	20	26,097
	RBBB	49,876	1,018	28	50,922
	LBBB	28,998	592	25	29,615
	SB	32,386	661	15	33,062
	AF	32,550	665	11	33,226
	ST	45,199	923	35	46,157
	Multi-label	22,203	454	12	22,669
	Total	**2,265,913**	**46,247**	**827**	**2,290,318**
Risk prediction	No AF	797,786	16,282	203,568	1,017,636
	Future AF	8,737	178	2,162	11,077
	Total	**806,523**	**16,460**	**205,730**	**1,028,713**
Age estimation	Regression	**1,852,130**	**115,759**	**233,233**	**2,312,160**

2) Physikalisch Technische Bundesanstalt Extra Large (PTB-XL) Dataset: The PTB-XL dataset [13] was used as a second dataset to verify the consistency of our experimental results. A total of 89 recordings without age available were excluded, leaving 21,748 recordings in total. For the arrhythmia diagnosis task, the same six classes as described for the TNMG dataset were used. The AF risk prediction task was not performed given that this dataset had no information about future AF development.

B. Feature Engineering

Three types of features were used. The first type consisted of 16 self-reported clinical variables (META; see supplementary

information Table SI). As for engineered features, a total of 23 heart rate variability (HRV) features were used, including those listed in Chocron et al. [14] as well as three "extended parabolic phase space mapping" features [15]. Moreover, a mean signal quality index called bSQI [16] was computed as an additional HRV feature (see Table SII for the elaborated details). The third set consisted of 22 morphological (MOR) features [17] (see Table SIII for an additional information). HRV and MOR features were computed for each lead, resulting in 557 engineered features per a 12-lead ECG recording. For the cardiac arrhythmia diagnosis task, from the META features, only age and sex features were available for patients included in the expert-reviewed test set. Therefore, the additional 14 META features were discarded, resulting in 543 features per recording example. All the features were used for the AF risk prediction task. For the age estimation task, 556 features were available, since age was used as the target label and thus not included in the feature set. For the PTB-XL dataset, only age and sex were available; thus, the arrhythmia diagnosis and age estimation tasks relied on 543 and 542 features, respectively.

C. Model Pipeline

The dataset split was different for each task (Table I). For the cardiac arrhythmia diagnosis task, data were stratified according to the six different arrhythmias and split to 98% train and 2% validation, as in the work of Ribeiro et al. [8]. The additional 827 recordings obtained by the expert consensus were used as the test set [8]. For the risk prediction task, the split was 80% train and 20% test, as in our previous work [11]. Then, 2% of the recordings taken from the train set split were used as the validation set. For the age estimation task, train and validation sets were split in a ratio of 80 to 5, as in Lima et al. [12], with the recordings taken from the CODE-15% [10] as the test set. The PTB-XL dataset split was made according to author recommendation [13], i.e., selecting their 9th and 10th pre-made folds as the validation and test sets, respectively. This ensured a stratified division according to the diagnosis task labels and kept the recordings of same patients in the same sets (i.e., no information leakage). Moreover, the validation and test folds had at least one human cardiologist evaluation, ensuring higher label quality [13].

Experiments for the three tasks were conducted with the same pipeline, and consisted of the following steps (Fig. 1(a)): (i) Setting the importance of the engineered features using the minimum redundancy maximum relevance (mRMR) algorithm [20]. (ii) Obtaining the FE performance score from a classical machine learning model by training a random forest (RF) classifier or regressor and taking FE as input. The number of selected features was set at this step to be later used as input to the concatenation layer (purple in Fig. 1(b)) in the merged FE+DL experiments. (iii) Hyperparameter tuning, including the fully connected (FC) layer adjustment added after the concatenation layer. (iv) Running the DL and FE+DL experiments by training the models in separated DL branch and merged FE+DL fashions to obtain their scores. The pipeline was applied in an end-to-end fashion for all experiments. Thus, there was no human intervention involved in the process.

1) Feature Importance With mRMR: The mRMR algorithm [20] from MATLAB R2020b (Mathworks) built-in functions was used. The mRMR was applied on the training set with respect to the relevant labels per task. For the diagnosis task, a multiclass-multilabel problem, all the different labels were considered as targets. Thus, we first treated the feature selection as multiple binary classification problems and applied mRMR with respect to each of the six labels (i.e., arrhythmias), obtaining six different rankings. Then, we used a union approach, i.e., when we set a number of selected features, there was at least that number of features suitable for each of the labels in the final unified selected features. For example, by selecting 352 features to use, we in practice selected all 543 features of the diagnosis task. This aligned with the approach used in the DL branch according to Ribeiro et al. [8], as classifying the six arrhythmias was executed using a sigmoid activation function in the final layer. Thus, it addressed the task as a multiclass-multilabel problem, and the probabilities were obtained for each label independently of the others. The mRMR was applied with respect to the labels of the other two tasks as well: future AF labels as target for the AF risk prediction task to rank the 557 features (binary classification), and patient age for the age estimation task to rank the 556 features (regression). See Fig. S1, S2, S3 for the mRMR results.

2) Model Training and Hyperparameter Tuning: RF classifier and regressor of the scikit-learn 1.0.2 [21] Python package were used to perform classification and regression tasks with the engineered features. Hyperparameter tuning was performed with respect to the validation set performance, while all final scores were reported on the test set. The optimized performance measure changed according to the task: harmonic mean of the positive predictive value (PPV) and sensitivity (Se), i.e., the F1-score, area under the receiver operating characteristic curve (AUROC) and negative mean absolute error (MAE), for the arrhythmia diagnosis, risk prediction, and age estimation tasks, respectively. The number of trees, maximum depth, split quality criterion and minimum number of samples at a leaf node were searched with a Bayesian optimization (scikit-optimize 0.9.0), iterating over a changing number of features (see Table SIV for the selected parameters). In the classification tasks, class weights were applied to compensate the class-imbalance in the training sets. The least number of features yielding a performance plateau over the validation score was selected for the later FE+DL experiment in each task. In this way, it was expected to reach a trade-off between the model degrees of freedom and its performance.

The neural network models were implemented using TensorFlow 2.5. The learning rate was set to 10^{-3} and was reduced by a factor of 10 when the validation performance stopped improving for five consecutive epochs, similarly to [8]. This allowed faster convergence in the initial epochs and finer refinements with smaller step sizes near an optimum. The final scores were reported on the separate test set. Hyperparameter tuning was performed with respect to the area under the precision recall curve (AUPRC) for the arrhythmia diagnosis task, with respect to the AUROC for the AF risk prediction task, and with respect to

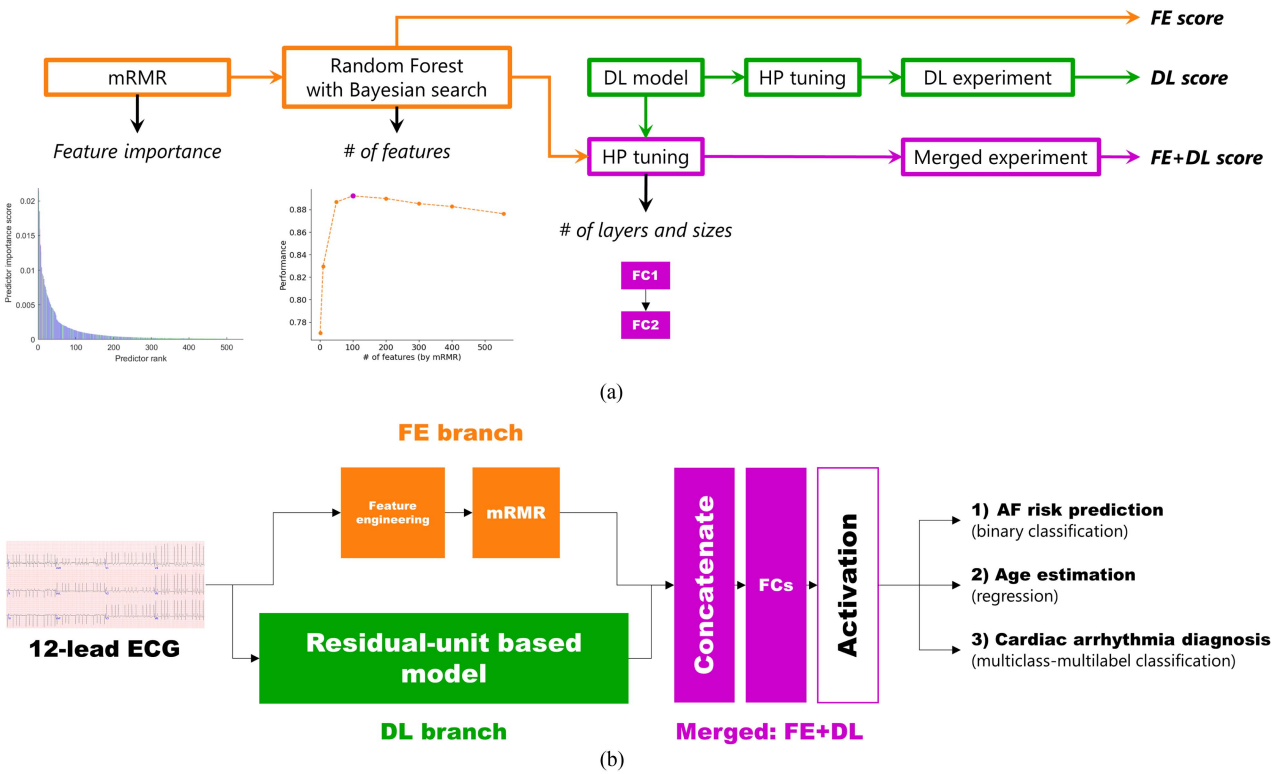

Fig. 1. Model pipeline and overall architecture. Orange, green and purple components are associated with the feature engineering (FE), deep learning (DL) and merged (FE+DL) experiments, respectively. (a) Pipeline steps and feature inclusion in the merged model: 1. Engineered features importance is set by minimum redundancy maximum relevance (mRMR) algorithm. 2. Random forest (RF) classifier or regressor is trained and optimized by Bayesian search with changing number of features prioritized by the first step results. Best number of features (based on the validation set) is used for next steps and best RF model performance (based on the test set) is taken as the FE score. 3. DL model hyperparameters (HP) are tuned. Then the DL experiment is conducted to obtain the DL score 4. HP tuning with merging the selected features and the DL model, including optimizing the fully connected (FC) layers after the FE and DL feature concatenation. 5. The merged experiment is conducted to obtain the FE+DL scores. (b) Overall model architecture design: a 12-lead ECG is the input for both branches. In the FE branch (orange), the signal is analyzed with feature engineering methods to extract features. Then, some of the features are selected with mRMR as explained in a. The DL branch (green) consists of deep neural network architecture taken from [8] (or [18] or [19]). Features from both branches are merged by concatenation in the purple head. These together are classified with FC layers and an activation that is altered according to one of the three output tasks.

the MAE for the age estimation task. Binary cross-entropy was used as the loss function for the cardiac arrhythmia diagnosis and AF risk prediction classification tasks. MAE was used as the loss function for the age estimation regression task. For all tasks, a batch size of 256 was used. Moreover, class-imbalance was corrected by assigning class weights according to their presence in the prediction and diagnosis task training sets. The number and size of the FC layers in the purple merging head (Fig. 1(b)) were tuned as hyperparameters.

3) DL Branch: The DL branch was based on the model reported by Ribeiro et al. [8], and later used by Lima et al. [12] and Biton et al. [11]. Thus, it was taken as a benchmark model for our 12-lead ECG data. Briefly, 4 residual blocks extract the deep features, and end with a final classifying FC layer. The residual blocks include 1D convolutional layers treating the input raw 12-lead ECG signal as a time-series of 12 channels. In addition to the baseline DL model we selected, i.e., from Ribeiro et al. [8], two additional state-of-the-art architectures developed by Han et al. [18] and Hannun et al. [19] were evaluated. This was performed for the purpose of ensuring that our conclusions were not dependent on the choice of a specific DL architecture. For the model by Han et al., the DL branch was taken starting from after the input layer and ending with the flatten layer. For the model from Hannun et al., we used the architecture in the code given by the author, starting from after the input layer and ending before the last dense layer. The models from Ribeiro and Hannun are similar in their structure, as they are constructed from residual blocks and are applied on the whole 12-lead ECG signal at once, whereas the Han model analyzes each lead individually before combining them. Moreover, the Ribeiro and Hannun models were originally designed for arrhythmia diagnosis, in contrast to the Han model which was created for detecting myocardial infarction.

4) Merged FE+DL Model Architecture: Fig. 1(b) shows the general model architecture including the FE branch (orange), the DL branch (green) and the merging head (purple) of the FE+DL experiments. The merging head included a concatenation layer which combined between the output neurons of both branches. These was followed by one or more FC layers to create the final network output. The final activation was task-based: sigmoid for classifications and no activation (linear) for the regression.

D. Statistical Analysis

1) Performance Measures: For the arrhythmia diagnosis task, the AUPRC and the F1-score were computed, as described in Ribeiro et al. [8]. In addition, we reported the corresponding false positive rate (FPR) and false negative rate (FNR). For the AF risk prediction task, the AUROC was used as in [11], [22]. Last, for the age estimation task, which is a regression problem, the coefficient of determination (R^2) was computed to characterize the linear fitting between the input and output ages as in [23], [12]. The MAE which served as a loss function, was reported as well to describe the estimation error. To compute the final AUPRC, the precision recall curve was processed by an interpolation and a median filter, followed by the trapezoid technique.

2) Confidence Intervals: For each task, the confidence intervals were evaluated for the three experiments (i.e., FE, DL, FE+DL) using bootstrapping as in [11]. Hence, the performances were computed on 1000 portions taken from the test set by randomly permuting its recordings and selecting only 80% in each iteration. Then, a student's t-test was performed on each of the three pairs (i.e., FE versus DL, FE versus FE+DL and DL versus FE+DL). The significance cut-off was defined as $p_{value} < 0.01$, which determined if the experiment results were significantly distinguishable.

E. Learning Curves

Learning curves were produced to investigate how the number of ECG recordings available for training affected the performance of the experiments. In particular, it enabled assessment of how many ECG recordings were needed for DL performance to reach FE performance. For all the three tasks, these learning curves were produced for all experiments (i.e., FE, DL and FE+DL). The original hyperparameter set found when using the full training set was also used to produce these curves.

III. Results

A. Cardiac Arrhythmia Diagnosis

FE performance was reported for each arrhythmia individually (Fig. 2(a)). All binary classification performances reached their plateau after selecting 50 features; therefore, for the FE+DL experiment, we included the 50 most important features with respect to each one of the labels. Since we worked with the union approach, this effectively involved 214 features to be used in the FE+DL experiment. For this task, there was no need for additional FC layers between the concatenation and the final layer.

The test set results for the three experiments are summarized in Table II. Best performance measures (i.e., highest AUPRC, highest F1-score, lowest FPR and lowest FNR) are in bold, showing overall comparable scores, with no clear preference for the FE, DL or FE+DL experiment. The validation set results are also reported in Table SV, and show no overfitting, as the test set results were better than the validation set for all the arrhythmias. All experiments obtained similar results, with

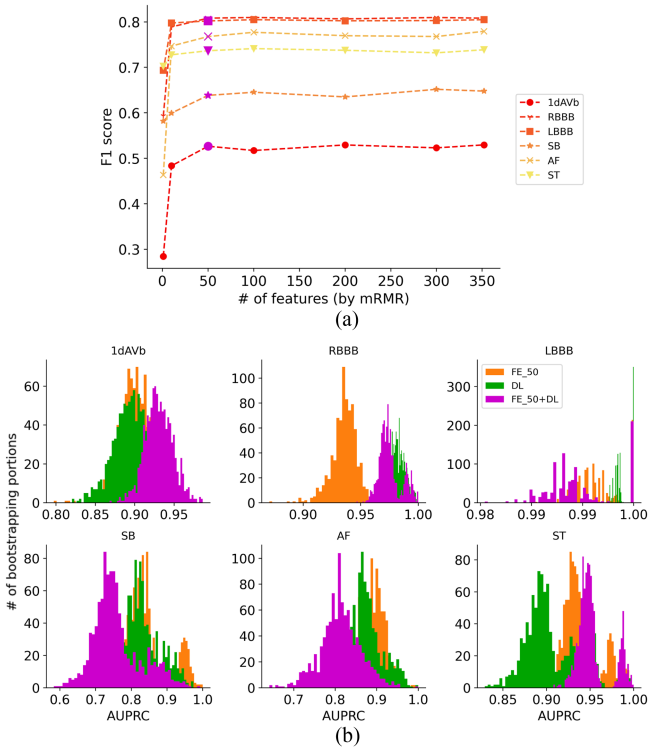

Fig. 2. Arrhythmia diagnosis experiments on the TNMG dataset. (a) Feature selection by minimum redundancy maximum relevance (mRMR) algorithm for the arrhythmia diagnosis task. Every line shows a separate random forest binary classification experiment based on different arrhythmia labels. The F1-score is reported on the validation set. 50 features (purple points) were selected for the feature engineering with deep learning (FE+DL) experiments, to include the features after all the classifiers reaching their plateau. (b) Bootstrapping over the test set results for each of the arrhythmias in each experiment. Orange, green and purple are the samples extracted from the test set results for the FE, DL and FE+DL experiments, respectively. The x-axis represents the area under the precision recall curve (AUPRC) score.

the FE mean AUPRC being the highest ($AUPRC_{FE} = 0.92$). Nevertheless, different experiments performed better than others with respect to specific arrhythmia diagnosis. For instance, for the RBBB arrhythmia, the DL experiment obtained best AUPRC performance ($AUPRC_{DL} = 0.98$), while, for AF, the FE experiment obtained highest AUPRC score ($AUPRC_{FE} = 0.89$). Moreover, AF was the only case when the FE+DL experiment failed to deliver similar performance to FE or DL scores. The arrhythmia precision recall curves in Fig. 3 show the classifier behavior for different Se-PPV trade-offs. The bootstrapping results over the test set are shown in Fig. 2(b). Although the performance scores were similar in most cases, all experiments yielded significantly different results, with $p_{value} \ll 0.01$, i.e., for FE versus DL, FE versus FE+DL and DL versus FE+DL.

B. Risk Prediction for AF

Fig. 4 shows the results for the AF risk prediction task. The optimized validation score was obtained using 100 of the 557 features (Fig. 4(a)), with importance set by the mRMR. With these, the RF classifier set a FE test score of $AUROC_{FE} = 0.86$

TABLE II
CARDIAC ARRHYTHMIA DIAGNOSIS TASK SCORES (TNMG)

Arrhythmia	Test set score	FE	DL	FE+DL
1dAVb	AUPRC	0.90	0.89	**0.93**
	F1	0.83	0.81	**0.88**
	FPR	0.008	0.009	**0.005**
	FNR	0.143	0.143	**0.107**
RBBB	AUPRC	0.94	**0.98**	0.97
	F1	0.91	**0.96**	**0.96**
	FPR	0.004	0.004	**0.003**
	FNR	0.088	**0.000**	0.029
LBBB	AUPRC	**1.00**	**1.00**	0.99
	F1	0.97	**0.98**	**0.98**
	FPR	0.003	**0.000**	**0.000**
	FNR	**0.000**	0.033	0.033
SB	AUPRC	**0.84**	0.83	0.75
	F1	0.83	**0.86**	**0.86**
	FPR	0.006	**0.005**	**0.005**
	FNR	**0.063**	**0.063**	**0.063**
AF	AUPRC	**0.89**	0.87	0.81
	F1	**0.82**	**0.82**	0.78
	FPR	**0.000**	**0.000**	0.001
	FNR	**0.308**	**0.308**	**0.308**
ST	AUPRC	0.94	0.91	**0.95**
	F1	**0.94**	0.93	**0.94**
	FPR	0.006	**0.004**	0.005
	FNR	**0.000**	0.054	0.027
Mean	AUPRC	**0.92**	0.91	0.90
	F1	0.88	0.89	**0.90**
	FPR	0.004	0.004	**0.003**
	FNR	0.100	0.100	**0.095**

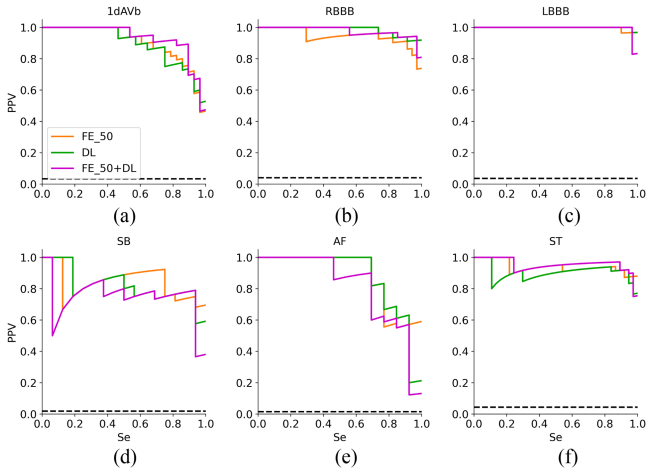

Fig. 3. Precision-recall curves, i.e., positive predictive value (PPV) versus sensitivity (Se), for the arrhythmia diagnosis task on the TNMG dataset. The results are presented for the test set. Orange, green and purple lines shows feature engineering (FE), deep learning (DL) and merged branches experiment of FE together with DL, respectively. The black line resembles performance for a no-skill classifier (all predictions are positive). The corresponding area under the precision-recall curve (AUPRC) scores are summarized in Table II.

(Fig. 4(b)). After training the DL branch alone, there was a small improvement to $AUROC_{DL} = 0.87$. For the FE+DL experiment, best performance for this task was obtained when adding another FC layer of 1000 neurons post-concatenation and before the final layer. The FE+DL experiment reached an AUROC score comparable to that of the FE ($AUROC_{FE+DL} = 0.86$). While

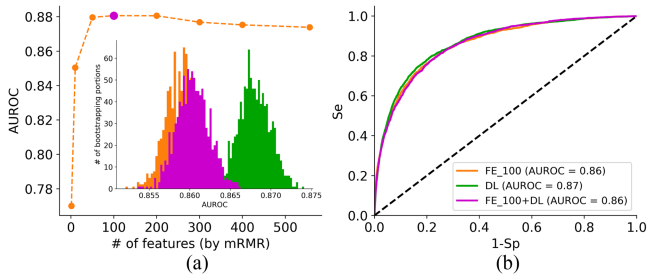

Fig. 4. AF risk prediction task results on the TNMG dataset. (a) Optimization of the number of features selected obtained by training a random forest (RF) classifier and reporting the area under the receiver operating characteristic curve (AUROC) performance over the validation set. Accordingly, 100 features were selected (purple point) as input to the merged experiments of feature engineering plus deep learning (FE+DL). Inset: Histograms after bootstrapping the three experiments (colors are according to the legend in (b)) for statistical comparison. (b) ROC curve, i.e., sensitivity (Se) versus one minus specificity (Sp), and AUROC scores for the three experiments computed on the test set.

bootstrapping over the test set, we observed that the DL performance increment was significant for both the FE and FE+DL experiments (Fig. 4(a) - inset); yet, the improvement was minor. The three AUROC curves (Fig. 4(b)) were almost superimposed, suggesting similar classification performance in all experiments. The validation set results are reported in Table SV, which shows no overfitting, as the results between the validation and test sets were comparable.

C. Age Estimation

For the age estimation task, the RF regressor reached a validation performance plateau already at 200 of the 556 features (Fig. 5(a)). With these, the regressor reached a test performance of $R^2_{FE} = 0.60$ and $MAE_{FE} = 10.64 + 7.6$ years (Fig. 5(b)). Here, we did not find it useful to add FC layers between the concatenation and the final layer. DL performed with $R^2_{DL} = 0.83$ and $MAE_{DL} = 6.32 + 5.38$ years, and DL+FE performed with $R^2_{FE+DL} = 0.83$ and $MAE_{FE+DL} = 6.26 + 5.35$ years (Fig. 5(c), (d)). The bootstrapping experiment in Fig. 5(a) shows a significant ($p_{value} \ll 0.01$) separation between DL and FE+DL, albeit with a very small fold-change in the mean performance. The results in Table SV, show no overfitting and comparable MAE performance between the validation and test sets.

D. Learning Curves

Fig. 6 shows model performance as a function of the training set size. For the arrhythmia diagnosis task (Fig. 6(a)), the performances of the FE and DL experiments were comparable for all different sizes, with a significant superiority for FE over DL data for small training sets ($< 500k$ examples). The FE+DL experiment demonstrated slower performance growth than the single models. For the AF risk prediction case (Fig. 6(b)), FE performed significantly better than DL for small training sets (up to $500k$ recordings). DL reached FE performance at approximately $700k$ recordings. Since the maximal training set size was $806k$

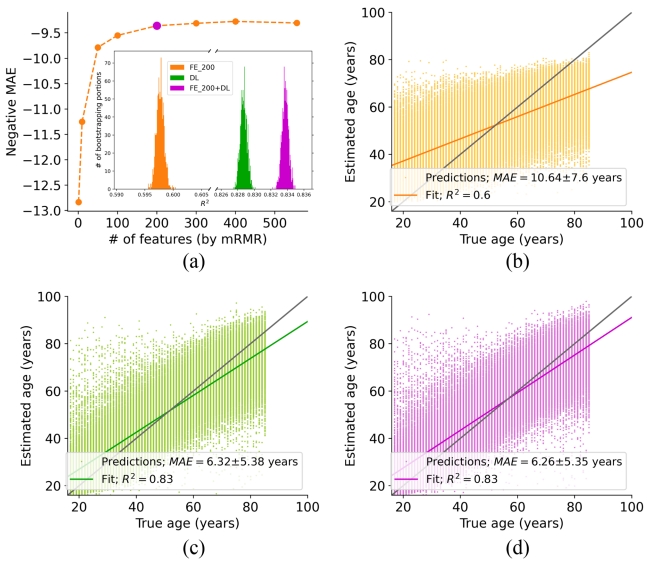

Fig. 5. Age estimation task results on the TNMG dataset. (a) Feature selection according to the random forest validation results. 200 features were selected, as performance reach plateau at this point (purple), according to the negative mean absolute error (MAE). Inset: Bootstrapping over the test set in the feature engineering (FE, orange), deep learning (DL, green) and FE+DL (purple) experiments. (b)–(d) Test set results for the regression task for the three experiments: FE, DL and FE+DL in orange, green and purple, respectively. The results for each recording are displayed with their MAE. Moreover, the overall linear fitting is displayed with the extracted R^2. The gray lines resemble identity between the estimated and true age ($y = x$).

recordings (Table I), it was not possible to determine whether DL continued to improve with more recording examples. For the age estimation task (Fig. 6(c)), DL consistently performed better than FE, with a clear advantage for all training set sizes $>$ 1k examples.

E. Comparison of Different DL Architectures

In addition to using the Ribeiro et al. model [8], we performed the main TNMG experiments for all tasks with the models of Han et al. [18] and Hannun et al. [19]. The Ribeiro and Hannun models achieved similar performances in all three tasks (Table SVI). For the arrhythmia diagnosis, the Hannun model obtained mean AUPRC scores of $AUPRC_{DL} = 0.93$ and $AUPRC_{FE+DL} = 0.89$ (compared to Ribeiro's $AUPRC_{DL} = 0.92$ and $AUPRC_{FE+DL} = 0.90$). For AF risk prediction, the AUROC scores obtained by Hannun were $AUROC_{DL} = 0.86$ and $AUROC_{FE+DL} = 0.85$ (compared to Ribeiro's $AUROC_{DL} = 0.87$ and $AUROC_{FE+DL} = 0.86$). Last, for age estimation, the R^2 scores obtained were $R^2_{DL} = 0.83$ and $R^2_{FE+DL} = 0.84$ (compared to Ribeiro's $R^2_{DL} = 0.83$ and $R^2_{FE+DL} = 0.83$). For the Han model, the scores were similar for the AF risk prediction task, with $AUPRC_{DL} = 0.87$ and $AUPRC_{FE+DL} = 0.87$, but lower for the arrhythmia diagnosis task ($AUROC_{DL} = 0.90$ and $AUROC_{FE+DL} = 0.80$), and in the age estimation task ($R^2_{DL} = 0.75$ and $R^2_{FE+DL} = 0.75$).

F. Experiments on the PTB-XL

In addition to the TNMG dataset, we performed the arrhythmia diagnosis and age estimation tasks using the PTB-XL dataset (see supplementary information Fig. S4, Table SVII, Fig. S5 and Fig. S6). For the arrhythmia diagnosis task, the selected number of features was 50 (Fig. S4), and the FE and DL experiments yielded comparable results, with mean AUPRC scores of $AUPRC_{FE} = 0.79$ and $AUPRC_{DL} = 0.77$ (Table SVII). Fig. S5 shows precision-recall curves for the different arrhythmias, where comparable results were achieved between FE and DL. The age estimation task results are presented in Fig. S6. A total of 300 features were selected to perform the age estimation task. These resulted in scores of $R^2_{FE} = 0.53$ and $MAE_{FE} = 9.84 + 7.39$ years for the FE experiment, versus $R^2_{DL} = 0.71$ and $MAE_{DL} = 7.49 + 5.74$ years for the DL experiment, and $R^2_{FE+DL} = 0.71$ and $MAE_{FE+DL} = 7.51 + 5.85$ years for the merged experiment, Thus, confirming the superiority of DL for this task.

IV. Discussion

For the cardiac arrhythmia diagnosis task, the FE and DL achieved comparable mean scores of $AUPRC_{FE} = 0.92$, $F1_{FE} = 0.88$ and $AUPRC_{DL} = 0.91$, $F1_{DL} = 0.89$. Thus, it can be debated what method is best, particularly when considering each individual arrhythmia. For instance, in the cases of RBBB, the DL experiment obtained a higher AUPRC score than FE ($AUPRC_{FE} = 0.94$ versus $AUPRC_{DL} = 0.98$), whereas for AF, the FE experiment yielded the higher score ($AUPRC_{FE} = 0.89$ versus $AUPRC_{DL} = 0.87$). The DL experiment reproduced the work of Ribeiro et al. [8], and reached a F1-score similar to those originally reported.

In terms of feature selection, we saw a consistency between the medical literature and the FE features prioritized by the mRMR (Fig. S2, S3). First, 1dAVb is associated with PR intervals greater than 200 ms [24], corresponding to the PR_{seg} feature (Table SIII) which was found to be the most important for this diagnosis (Fig. S2). Both RBBB and LBBB are characterized by prolonged QRS complexes and modified R-wave [25], [26]. This is in accordance with several of the most important features found: QRS_{int} and $Rwave$ for RBBB and LBBB, and QRS_{Area} for LBBB. Moreover, for the RBBB, the $Jpoint$ (Table SIII), which is associated with the QRS complex [27], was flagged as the most important feature. For SB and ST, which are associated with an abnormally low and high rhythm, respectively [28], a large number of HRV features were selected (Fig. S3). In particular, the median heart rate (medHR) was the most important feature in both arrhythmias. Finally, AF diagnosis relies on variations and the absence of the P-wave [29]. Accordingly, the MOR features PR_{seg} and $Pwave$ which can point to the missing P-wave were two out of the three most important features, where the first one was the AFEv HRV feature (Table SII), developed specifically to diagnose AF [30].

For the AF risk prediction task, the results of the three models were comparable (Fig. 4(b)), with DL yielding the best AUROC score of 0.87. In our previous work [11], we predicted AF by

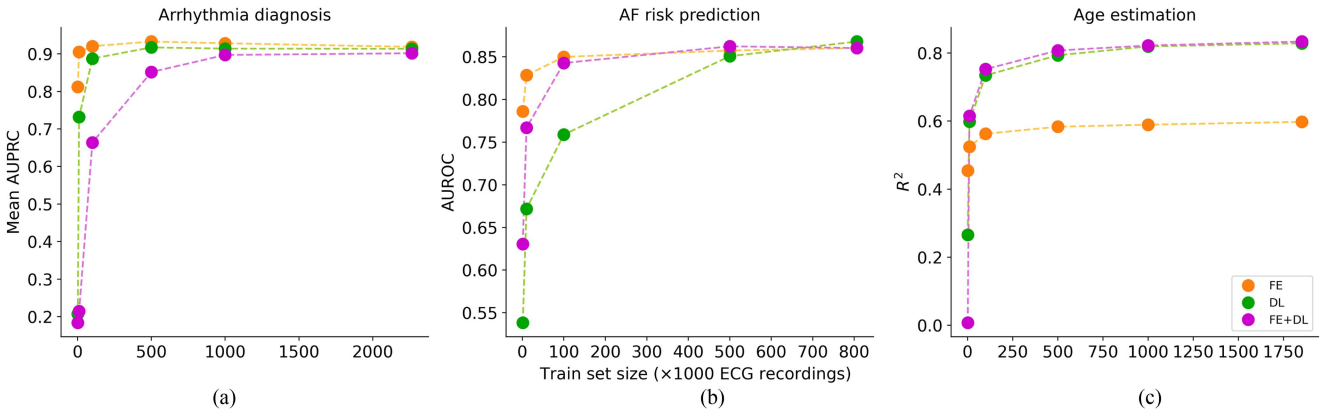

Fig. 6. Learning curves. Altering train set size experiments for the three different tasks on the TNMG dataset: (a) arrhythmia diagnosis; (b) atrial fibrillation (AF) risk prediction; and (c) age estimation. For all the tasks, smaller train set size experiments were added including $1k$, $10k$, $100k$, $500k$ or $1M$ ECG recordings, while the figures also show the maximum train set size that was used in the main experiments (Table I). All the different experiments - feature engineering (FE, orange), deep learning (DL, green) and merged (FE+DL, purple) were conducted. The performance measures used here aligned with the main experiments: (a) the mean of the area under the precision recall curve (AUPRC) computed over the six arrhythmias; (b) area under the receiver operating characteristic curve (AUROC); and (c) the coefficient of determination (R^2). The horizontal axis is in units of thousand ECG recordings.

exploiting a hybrid method, extracting latent features from a deep neural network which was previously trained for detecting AF presence. These representation learning features were used as input to a RF classifier. Using the same dataset as in the present experiment, this network achieved a score of $AUROC = 0.82$, and thus was outperformed by the DL experiment conducted here. Therefore, our research demonstrated how the end-to-end DL method was a better option for this task. Nevertheless, when considering integration of FE, Biton et al. showed an AUROC of 0.91, which outperforms the current results. The higher performance of this hybrid model [11] was due to the pretraining of the deep neural network on the additional $> 1M$ recording examples that included AF examples, suggesting that this pretrained network had a significantly enhancing impact on the performance. Thus, if pretraining is possible, it may represent an advantage of DL over FE.

Regarding feature importance, the best 10 features selected for this task by mRMR included PR_{seg}, $Pwave$ and $Pwave_{int}$ (Fig. S1, Table SIII). AF prediction is known to be associated with PR-interval duration, P-wave duration and the absence of P-wave amplitude [31], [32], [33], [34]. Thus, the selected features correspond to those found in research. The IrrEv and PACEv features (Table SII) are included as well between the best feature predictors. These were developed to detect AF and were therefore also associated with the specific arrhythmia [30]. Finally, for this task, we saw more contributing META features, specifically the age feature which was ranked as the best AF predictor, which is consistent with the high prevalence of AF in advanced ages [35].

For the age estimation task, DL significantly outperformed FE (Fig. 5(b) versus Fig. 5(c)). Only the DL was adequate to extract the necessary features relating to patient age. This may be explained by the fact that engineered features are inherently drawn from domain knowledge about the relation between HRV or MOR features and cardiac conditions, but not about patient demographics such as age. This suggests that for tasks with very little domain knowledge available, DL may be a better approach. Furthermore, it achieved a $R^2 = 0.83$, which was higher than in previous reports by Attia et al. [23] and Lima et al. [12] ($R^2 = 0.7$ and 0.71, respectively). The difference from Attia et al. [23] may be explained by the larger dataset used in our experiments ($> 2.3M$ versus $> 770k$ recordings). Therefore, our experiment leveraged the potential of DL to improve performance when trained on a large dataset (Fig. 6(c)). In Lima et al. [12], the experiments were almost identical with respect to the data used. Nevertheless, the authors' model was originally tuned to predict mortality and thus included proportional weighting of the train set age groups to increase performance towards this specific target. This strategy proved to be efficient in transferring the age estimation model to the mortality prediction task, but led to reduced performance for the age estimation task itself.

In Fig. S1, the three top-ranking age estimation features were $Rwave$ (MOR, Table SIII), RMSSD (HRV, Table SII) and calcium blockers (META, Table SI). Other important META features such as COPD, diuretics and smoking suggest that aging is associated with an increased prevalence of diseases and health risking factors.

For the three tasks, learning curves were produced in order to obtain insights into the minimal number of recordings necessary to reach best performance. For the arrhythmia diagnosis task, the learning curves showed that in a small data regime, FE outperformed DL and that in a big data regime, there was no significant advantage for FE or DL (Fig. 6(a)). For the AF risk prediction task, we also noted that FE outperformed DL in a small data regime. However, DL started to outperform FE at about $700k$ recordings (Fig. 6(b)). There, the prediction task was more complex and less directly related to the original FE domain knowledge (as opposed to diagnosis), which possibly enabled DL to outperform FE as the training dataset became larger. For the age estimation task, the FE and the DL curves were distant from one another from very few recording examples and onward (Fig. 6(c)), with DL consistently outperforming FE. We explain this by the fact that the age estimation task is

far from the historical clinical tasks that traditional FE were designed for. Importantly, the altering train set size experiments demonstrated how certain tasks would actually require a relatively small dataset size to reach maximum performance. Specifically, the arrhythmia diagnosis could reach the maximal mean performance by using only $100k$ ECG recordings available with the FE approach, whereas a comparable performance with DL would have necessitated $500k$ recordings. For the AF risk prediction task, we could achieve with FE a high performance of an $AUROC = 0.85$ by using only $100k$ ECGs, while DL needed seven times more data to match this score. Finally, the age estimation task reached a plateau after $500k$ recordings, regardless of the approach. Overall, these learning curves provide valuable insights into the advantages or lack thereof, of DL over FE and the amount of data that is necessary to reach best performance.

Overall, the results showed that FE and DL performances were comparable for the two clinical tasks (diagnosis and risk prediction), while DL significantly outperformed FE ($R^2_{DL} = 0.83$ versus $R^2_{FE} = 0.60$) for the age estimation task. The HRV and MOR features were historically often engineered for the purpose of supporting diagnosis and were built according to medical scientific knowledge (domain knowledge) to serve this purpose. In that respect, we explain the comparable performance of FE and DL for such tasks, while DL would perform better than FE for nontraditional clinical tasks such as age estimation from the ECG data.

PTB-XL experiments: The arrhythmia diagnosis and age estimation experiments were repeated on PTB-XL and led to similar findings. Specifically, the performances of FE and DL for the arrhythmia diagnosis task were comparable (mean AUPRC score of $AUPRC_{FE} = 0.79$ versus $AUPRC_{DL} = 0.77$), while better performance was obtained using DL versus FE for the age estimation task ($R^2_{FE} = 0.53$ versus $R^2_{DL} = 0.71$). Since the TNMG and PTB-XL datasets are from different medical institutions and geographical regions, these results strengthen our findings.

Effects of the DL architecture: The comparison between the three state-of-the-art models (Ribeiro et al. [8], Han et al. [18] and Hannun et al. [19]) in performing the three tasks demonstrated that our findings were not dependent on the specific choice of the Ribeiro DL architecture. Indeed, all models performed in a comparable way with the risk prediction task, whereas the model from Han et al. performed worse than the other for the arrhythmia diagnosis and the age estimation tasks.

Limitations: One important limitation was the limited test set size for the arrhythmia diagnosis task. The test set included only a small number of recordings for each arrhythmia that were reviewed by a board of cardiologists (Table I). This limited the ability to assess the relative performance of each approach (FE, DL, FE+DL), and to determine whether the fluctuations observed for each arrhythmia class would hold. Furthermore, despite the very large size of the TNMG dataset used for our experiments, it is important to note that the number of cases for each arrhythmia was moderately high (Table I: 25,555 - 49,876 recordings). Thus, it will be of value to further increase the dataset by adding new arrhythmia recordings. Another important limitation was the number of recordings used in the train set of the AF risk prediction task, i.e., $806k$ recording examples against at least $1.8M$ for the other tasks, which prevented us from further investigating whether the DL performance would continue increasing with more data (Fig. 6(b)). An additional limitation was the explainability of the deep features. A key aspect in understanding the gap between FE and DL lies in the comparison and explanation of the information encoded in the deep features. While this extended beyond the scope of the present research, this should be further investigated.

V. Conclusion

We developed a unified model pipeline and trained it for three different and independent cardio-related tasks, utilizing both feature engineering and deep learning approaches as well as combining both. Experiments were conducted using a large dataset of over 2.3M 12-lead ECG recordings. For all tasks, cardiac arrhythmia diagnosis, atrial fibrillation risk prediction and age estimation, the DL model reached similar or higher performance than the FE approach. However, DL improvement over FE was important only for the age estimation task. We explain this finding by the fact that age estimation is not a classical clinical task that the 12-lead ECG was historically used for. Consequently, FE encompassing domain knowledge for this specific task was low. For the classification tasks, it was found to be difficult to justify a big data regime and the complexity of a DL model, since the classical machine learning approach utilizing a much smaller dataset reached similar performances.

V. Acknowledgment

We thank S. Biton and S. Gendelman for their assistance with feature engineering. EZ acknowledges The Miriam and Aaron Gutwirth Memorial Fellowship.

References

[1] A. Mincholé et al., "Machine learning in the electrocardiogram," *J. Electrocardiol.*, vol. 57, pp. S61–S64, Nov. 2019.
[2] S. Sahoo et al., "Machine learning approach to detect cardiac arrhythmias in ECG signals: A survey," *IRBM*, vol. 41, no. 4, pp. 185–194, Aug. 2020.
[3] S. Hong et al., "Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review," *Comput. Biol. Med.*, vol. 122, 2020, Art. no. 103801.
[4] A. Voulodimos et al., "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–13, 2018.
[5] A. Bogatskiy et al., "Lorentz group equivariant neural network for particle physics," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 969–979.
[6] Z. I. Attia, G. Lerman, and P. A. Friedman, "Deep neural networks learn by using human-selected electrocardiogram features and novel features," *Eur. Heart J.- Digit. Health*, vol. 2, no. 3, pp. 446–455, 2021.
[7] T. Beer et al., "Using deep networks for scientific discovery in physiological signals," in *Proc. Mach. Learn. Healthcare Conf.*, 2020, pp. 685–709.
[8] A. H. Ribeiro et al., "Automatic diagnosis of the 12-lead ECG using a deep neural network," *Nature Commun.*, vol. 11, no. 1, Dec. 2020, Art. no. 1760.
[9] A. L. P. Ribeiro et al., "Tele-electrocardiography and bigdata: The CODE (Clinical Outcomes in Digital Electrocardiography) study," *J. Electrocardiol.*, vol. 57, pp. S75–S78, Nov. 2019.
[10] A. H. Ribeiro et al., "CODE-15%: A large scale annotated dataset of 12-lead ECGs," Jun. 9, 2021, doi: 10.5281/zenodo.4916206.
[11] S. Biton et al., "Atrial fibrillation risk prediction from the 12-lead electrocardiogram using digital biomarkers and deep representation learning," *Eur. Heart J. - Digit. Health*, vol. 2, no. 4, pp. 576–585, 2021.

[12] E. M. Lima et al., "Deep neural network-estimated electrocardiographic age as a mortality predictor," *Nature Commun.*, vol. 12, no. 1, 2021, Art. no. 5117.
[13] P. Wagner et al., "PTB-XL, a large publicly available electrocardiography dataset," *Sci. Data*, vol. 7, no. 1, Dec. 2020, Art. no. 154.
[14] A. Chocron et al., "Remote atrial fibrillation burden estimation using deep recurrent neural network," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 8, pp. 2447–2455, Aug. 2021.
[15] S. Moharreri et al., "Extended parabolic phase space mapping (EPPSM): Novel quadratic function for representation of heart rate variability signal," in *Proc. Comput. Cardiol.*2014, pp. 417–420.
[16] J. Behar et al., "ECG signal quality during arrhythmia and its application to false alarm reduction," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 6, pp. 1660–1666, Jun. 2013.
[17] S. Gendelman et al., "PhysioZoo ECG: Digital electrocardiography biomarkers to assess cardiac conduction," in *Proc. Comput. Cardiol.*, 2021, pp. 1–4.
[18] C. Han and L. Shi, "ML–ResNet: A novel network to detect and locate myocardial infarction using 12 leads ECG," *Comput. Methods Programs Biomed.*, vol. 185, Mar. 2020, Art. no. 105138.
[19] A. Y. Hannun et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Med.*, vol. 25, no. 1, pp. 65–69, Jan. 2019.
[20] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinf. Comput. Biol.*, vol. 3, no. 2, pp. 185–205, Apr. 2005.
[21] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
[22] A. H. Kashou et al., "A comprehensive artificial intelligence–enabled electrocardiogram interpretation program," *Cardiovasc. Digit. Health J.*, vol. 1, no. 2, pp. 62–70, Sep. 2020.
[23] Z. I. Attia et al., "Age and sex estimation using artificial intelligence from standard 12-lead ECGs," *Circulation: Arrhythmia Electrophysiol.*, vol. 12, no. 9, pp. 1–11, Sep. 2019.
[24] S. Cheng et al., "Long-term outcomes in individuals with prolonged PR interval or first-degree atrioventricular block," *JAMA*, vol. 301, no. 24, pp. 2571–2577, Jun. 2009.
[25] D. Da Costa, W. J. Brady, and J. Edhouse, "Bradycardias and atrioventricular conduction block," *BMJ*, vol. 324, no. 7336, pp. 535–538, Mar. 2002.
[26] P. Francia et al., "Left bundle-branch block–pathophysiology, prognosis, and clinical management," *Clin. Cardiol.*, vol. 30, no. 3, pp. 110–115, Mar. 2007.
[27] J. E. Hollander et al., "Standardized reporting guidelines for studies evaluating risk stratification of emergency department patients with potential acute coronary syndromes," *Ann. Emerg. Med.*, vol. 44, no. 6, pp. 589–598, Dec. 2004.
[28] D. H. Spodick, "Normal sinus heart rate: Sinus tachycardia and sinus bradycardia redefined," *Amer. Heart J.*, vol. 124, no. 4, pp. 1119–1121, Oct. 1992.
[29] F. Censi et al., "P-wave variability and atrial fibrillation," *Sci. Rep.*, vol. 6, May 2016, Art. no. 26799.
[30] S. Sarkar, D. Ritscher, and R. Mehra, "A Detector for a chronic implantable atrial tachyarrhythmia monitor," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 3, pp. 1219–1224, Mar. 2008.
[31] J. B. Nielsen et al., "Risk of atrial fibrillation as a function of the electrocardiographic PR interval: Results from the Copenhagen ECG Study," *Heart Rhythm*, vol. 10, no. 9, pp. 1249–1256, Sep. 2013.
[32] S. Bidstrup et al., "Role of PR-Interval in predicting the occurrence of atrial fibrillation," *J. Atrial Fibrillation*, vol. 6, no. 4, pp. 90–94, Dec. 2013.
[33] T. Yoshizawa et al., "Prediction of new onset atrial fibrillation through P wave analysis in 12 lead ECG," *Int. Heart J.*, vol. 55, no. 5, pp. 422–427, 2014.
[34] G. Conte et al., "Usefulness of P-Wave duration and morphologic variability to identify patients prone to paroxysmal atrial fibrillation," *Amer. J. Cardiol.*, vol. 119, no. 2, pp. 275–279, 2017.
[35] P. Kirchhof et al., "2016 ESC guidelines for the management of atrial fibrillation developed in collaboration with EACTS," *Eur. Heart J.*, vol. 37, no. 38, pp. 2893–2962, Oct. 2016.