ECG Segmentation by Neural Networks: Errors and Correction

1st Iana Sereda
dept. of Control Theory and Dynamics
Nizhny Novgorod State University
Nizhny Novgorod, Russia
sereda@itmm.unn.ru

2nd Sergey Alekseev
dept. of Control Theory and Dynamics
Nizhny Novgorod State University
Nizhny Novgorod, Russia
sergey.alekseev@itlab.unn.ru

3rd Aleksandra Koneva
dept. of Control Theory and Dynamics
Nizhny Novgorod State University
Nizhny Novgorod, Russia
aleksandra.koneva@itlab.unn.ru

4th Roman Kataev
dept. of Control Theory and Dynamics
Nizhny Novgorod State University
Nizhny Novgorod, Russia
r.kataev.unn@gmail.com

5th Grigory Osipov
dept. of Control Theory and Dynamics
Nizhny Novgorod State University
Nizhny Novgorod, Russia
osipov@vmk.unn.ru

***Abstract*—In this study we examined the question of how error correction occurs in an ensemble of deep convolutional networks trained for an important applied problem: segmentation of Electrocardiograms(ECG). We also explore the possibility of using the information about ensemble errors to evaluate a quality of data representation built by the network.**

***Index Terms*—convolutional neural networks, cardiac cycle, segmentation, ensemble, outliers, errors**

I. Introduction

Correction of errors of Artificial Intelligence (AI) systems is recognized as one of the main problems in the AI-based technical revolution [1]. The effect of error correction often appears in ensembles of neural networks. It is known that in most cases an ensemble performance is better than performance of any individual network in it [2]. The creation of an ensemble of models is widely used in modern machine learning as the last step of the working pipeline. However, it is difficult to predict which mistakes the ensemble can eliminate from the basic model errors and which cannot.

This problem of possible mistakes of the trained model remains relevant because the internal representation in the neural network is difficult to interpret [3]. The reliability of a neural network is directly connected to the quality of the internal representation which it has built. In the context of medical tasks, the problem of analyzing the quality of representation (and fixing the flows in it) is especially important due to the peculiarities of medical datasets: some pathologies are often represented by a small number of samples, while variants close to the normal may occur too often [4]. However, these pathological cases are most important because pathological morphologies of the cardiac cycle are strong manifestations of diseases.

Imbalance of classes in the data set often leads to the situation there formal quality metrics can give unreasonably good result, while the network could not cover all important aspects of data well. It is not always possible to correct data imbalances with well-known methods (such as, for example, oversampling), because is not always clear which particular classes require balancing. We illustrate the above problem using the example of the ECG segmentation task: all meaningful components of cardiac cycle (P-wave, T-wave and the QRS complex) are roughly balanced in most of ECGs due to the periodicity of ECG structure. But the task itself contains imbalance because the dataset is not balanced for diseases. Diseases change the morphology of the components of the cardiac cycle in different ways, so the representation built by the neural network must contain information of how the cardiac cycle looks like for every pathology presented in the dataset. When creating a quality metric for an arbitrary task, it is difficult to take into account the imbalance for all the hidden factors of influence for this task. In this paper we use data on how exactly the ensemble of networks corrects the errors of the single network in order to conclude about the quality of internal representation in the network.

One of the ways to investigate the reliability of the internal representation received within the network is to use adversarial examples [5]. Another common approach to analyse the quality of the representation of deep networks is based on the visualization of the learned attributes of different levels [6]. For models with attention, attention visualization can be used [7]. A new direction is to find a metric for evaluating the degree of disentanglement in representations [8]. Other methods can be found in a survey [3].

This paper is organized as follows. Sections II,III and V describe the applied task, sections IV, VI and VII describe the convolutional network, its training and ensemble training, in section VIII we give a qualitative analysis of the factors that influence error correction by the ensemble and demonstrate some interesting effects arising in the training of individual networks in the ensemble. Sections IX and X summarize the

The study was supported by the Ministry of Education and Science of Russia (Project No. 14.Y26.31.0022).

978-1-7281-2009-6/$31.00

main results.

The projects code is publically available at: https://github.com/Folifolo/SegmentationECG

II. Dataset

LUDB [9] is an open access dataset, containing ECG recordings of 200 unique patients. Each recording is represented by a 10-second signal registered from twelve leads with a sampling rate of 500 Hz. An expert's annotation is provided for each patient, annotating the three segments of the cardiac cycle: P, QRS and T. The proper detection of these waves/complexes is essential for ECG-based diagnostics of the cardiovascular system. A schematic representation of the cardiac cycle is shown in fig. 1. A significant part of the dataset

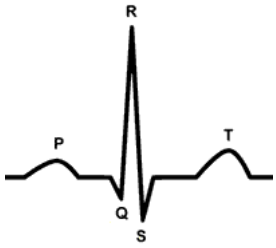


Fig. 1: Schematic cardiac cycle

is represented by healthy cases and the remaining part covers a wide range of different pathologies of the cardiovascular system (different heart rhythm types, conduction abnormalities, repolarization abnormalities and so on). The dataset also contains ECGs with varying degree of noise.

III. Data preparation

ECG preprocessing has consisted of Baseline wander (BW) removal, which is a conventional first step in ECG processing for most applications [10]. Baseline wander is a low-frequency ECG artifact, which may be caused by patients breath or movement [11] and holds no diagnostic information. ECG was filtered with two median filters as described in [10]. The resulting ECG is shown in fig. 2. High frequency noise was not removed, and no augmentation was performed.

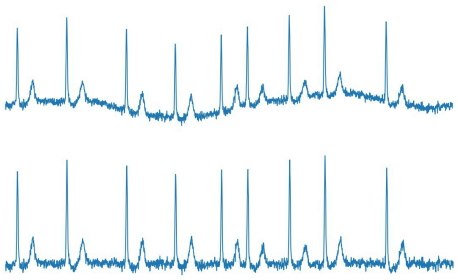


Fig. 2: BW removal example. ECG signal before and after processing is shown at the top and bottom panels respectively

IV. Base neural network

It was shown that the convolutional architecture of neural networks is well-suited for many real world problems. In particular, convolution networks have found application in medical data processing, including ECG analysis [12]. Therefore the architecture of the base network of the ensemble was chosen to be convolutional(shown in fig. 3).

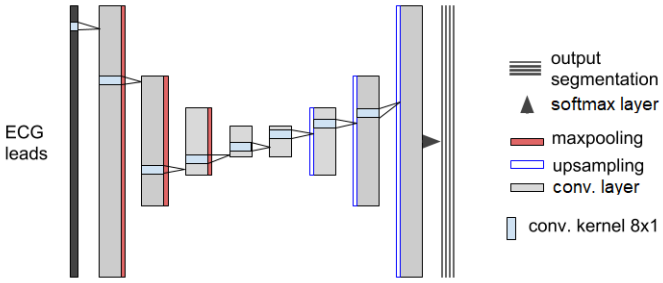


Fig. 3: An 8-layer convolutional architecture used for all experiments below

The softmax layer provides a probability distribution over 4 possible options for a given point in time: refer it as part of the T peak, the P peak, the QRS complex, or none of the three listed.

The final result represents binary masks for all three types of peaks. For each channel, the mask is generated based on the softmax output: for each point in time, a winner channel is selected. It gains 1, the remaining channels get 0. Examples of input ECG, corresponding softmax output and resulting delineation are shown in fig. 6 and 7.

Training was conducted on 12 ECG leads. However, a decrease in the number of leads under consideration does not greatly impair the result.

V. Quality evaluation

In this work we define each peak (complex) as the first and the last points of it. To evaluate the annotation quality for a particular type of points, (such as the P-wave starting points) we employ an algorithm that works as follows: for each point of this type on the ground truth annotation the algorithm looks for the corresponding point on the network's annotation.

If a corresponding point is found in the specified neighborhood of the point in ground truth annotation, then we count the networks decision as valid (True Positive, TP). In this case the error value is calculated as the distance between the point in ground truth annotation and the corresponding point in the network's annotation.

If a point specified by the network does not exist on the ground truth annotation in the specified neighborhood, we count the answer as false positive (FP). Should the network be unable to locate a point, which is present in the ground truth annotation, then the answer is marked as false negative (FN).

The permitted neighborhood is calculated adaptively depending on the patient's heart rate. For a heart rate of 70 BPM

the radius of the permitted neighborhood was chosen to be 150 ms. Then, the size of this neighborhood is decreased linearly based on the length of the cardiac cycle. An interval of 150 milliseconds was selected in accordance with ANSI/AAMI-EC57:1998 [13].

The following quality metrics are commonly used for ECG annotation evaluation:

- m – expected value of error
- σ^2 – error variance
- $Se = \frac{TP}{TP+FN}$ – sensitivity
- $PPV = \frac{TP}{TP+FP}$ – positive predictive value

Values of these metrics for the single network can be seen in table I and are slightly worse than that of the best direct methods [9]. However, it must be said that these metrics do not account for the degree of representation of various pathologies (diseases) within the dataset for which ECG segmentation is performed. For example, if the majority of samples belongs to healthy patients and if the segmentation algorithm handles the standard healthy case right(but not the pathological one) then its quality assessment will directly depend on the number of pathological or artifact-containing samples in the data set.

So it is important to investigate the behavior of the network on pathological cases rather than relying on formal metrics.

VI. Single network trends

In this section we describe some qualities that the single deep network demonstrates while being trained throughout the training dataset. The dataset was split into training (134 patients) and test (66 patients) parts. We did not single out the validation part due to the peculiarities of the dataset, namely due to its small size in combination with a high degree of imbalance. The reliability of the results is ensured by multiple runs of the experiment with different random partitioning into train and test.

Mini-batches are formed out of randomly selected 6-second intervals. The chosen optimizer is RMSProp [14] and the loss function is categorical cross entropy.

A. Noise stability

All else being equal, the presence of high-frequency noise (power line interference, 50 Hz) does not have a significant effect on the networks ability to produce a correct annotation, nor does it influence the smoothness of its output signal.

Absence of necessity of noise filtration is an advantage of the deep learning approach, since it reduces the time spent on ECG preprocessing.

B. Reaction to pathologies

It turned out that the presence of pathology(diseases) has the most noticeable effect on the quality of the neural network's performance.

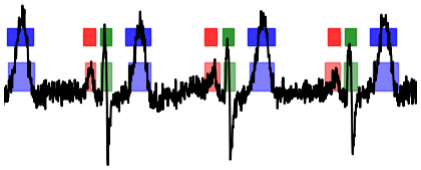


Fig. 4: Noised ECG is annotated correctly by a single network. The bottom row of colored boxes shows the annotation given by an expert, the top row depicts the annotation by the neural network, which architecture is depicted in fig. 3

When analyzing the network's performance on a test set of patients, it turned out that the following rule of thumb is valid: if the case is pathological, it can still be properly annotated by a single neural network. For example fig.5 shows an ECG with a non-standard T-wave shape and the network had annotated it well.

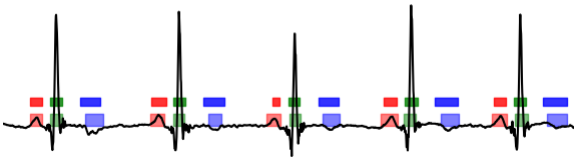


Fig. 5: An abnormal case (containing unusual T-waves) is annotated by a single network with satisfactory results

But if the annotation from the neural network was essentially erroneous (F1-score less than 0.9), then this ECG is pathological (or unreadable due to artifacts of the recording process).

This is especially true for the QRS complex. The QRS complex turned out to be the easiest to mark up with a neural network (as well as direct algorithms). Typically it has the largest signal amplitude although this is not always true, for example see fig. 4.

If we study the raw(softmax) output of the network for an ECG which does not contain any noticeable pathologies and compare it against the output for a markedly pathological ECG, we can notice a systematic difference in a raw signal. In the pathological case the networks signal contains numerous asymmetric low-amplitude surges, which do not contribute to the resulting annotation (see fig.7). However, we do not observe this kind of behavior in the non-pathological case: if the channel contain a signal surge, it is smooth and has a large enough amplitude to influence the annotation (fig. 6). In a sense, the intensity of this effect can be interpreted as the network's "confidence".

VII. Ensemble

To date, there are many strategies for the formation of neural network ensembles (for review see [15]). The selected strategy determines the number of members of the ensemble, the way they are combined and the way to guarantee diversity among them (for example, different learning rate protocols for the networks [2], correlation penalty [16], etc.). In our case,

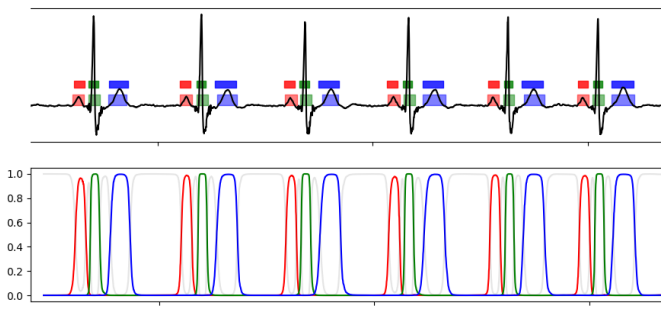


Fig. 6: A simple case of ECG annotation by a single network. On the top graph, the bottom set of colored markings shows the ground truth manual annotation for P waves (red), QRS complexes (green) and T waves (blue). The set of colored markings above represents the networks annotation for the same ECG. The bottom graph shows the raw output signal of the network. These values represent the networks confidence in the fact that the current segment does contain the appropriate ECG waveform. For a simple case of a healthy ECG, we can see that the network performs well when compared against a professional annotation done by a cardiologist. Smooth symmetric waves in the output signal of the network are characteristic of the segmentation of simple cases (like this one)

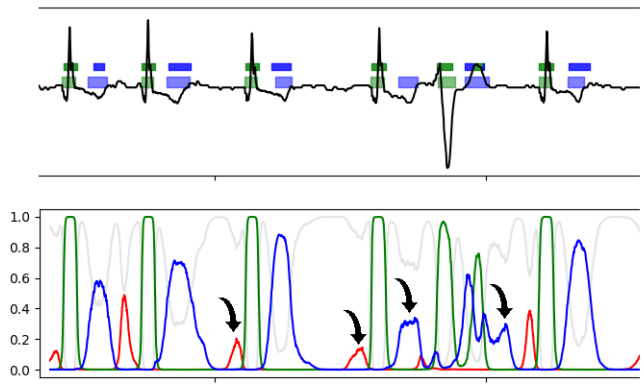


Fig. 7: Distinctive features of the networks output signal in a markedly pathological case. The top graph contains annotations for QRS complexes and T-waves in green and blue respectively. The bottom set of markings represents the ground truth annotation, the networks annotation is located above. The P wave (red) is lacking completely, a fact which was both noted by the cardiologist and the network. Non-smooth asymmetric waves of arbitrary amplitude in the output signal of the network are characteristic of segmentation of an electrocardiogram with severe pathology (like in this case) and are shown by black arrows

the diversity was provided by different training samples for members of the ensemble. The ensemble formation procedure was designed in such a way that adding a new network to the ensemble fixes some errors of the already existing networks on the training set.

After training the F1-score of the base network was measured on each patient of the training sample to see in which cases the network does not perform well. Then the procedure of iterative ensemble building starts. All patients rated at an F1-score of 0.99 and above were removed from the training set. The rest of the training set is then used as a separate training set for the new neural network.

This new neural network is then trained on that training set and, again, the procedure of screening out patients is carried out. All the cases on which this neural network has failed to achieve a score of 0.99 remain while others are deleted. The procedure described is repeated until the patients in the training set run out.

At each iteration of this algorithm a new neural network is created. Each of these networks is trained on an ever decreasing data set. If, after one step, the sample size has not changed then the same network is re-trained on the same sample on the assumption that it fell into a bad local minimum.

Figure 9 demonstrates an example of how the size of the training set has changed during the procedure described.

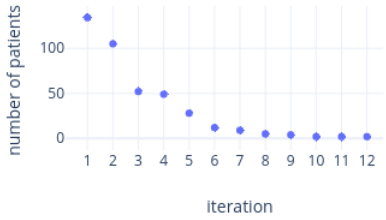


Fig. 8: Number of patients remaining in the train subset at every stage of ensemble formation. The formation continues until the number of patients in the subset reaches zero

When the ensemble is created, the resulting annotation for every input ECG can be obtained by averaging softmax output signals across all members of the ensemble.

VIII. Error correction in action

From the very ensemble construction algorithm itself one can see that the ensemble is able to systematically correct some errors of a single network. Members of the ensemble can correct each other's mistakes. To illustrate that, we provide a couple of typical examples. Fig. 9 demonstrates how the 4th member of the ensemble fixes the error of the 3rd member in case of an abnormal ECG.

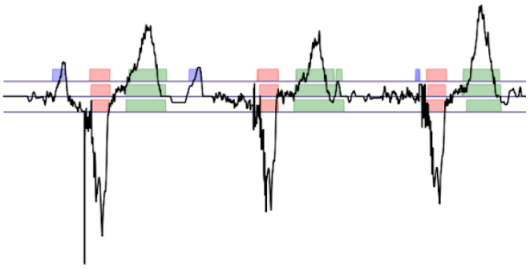


Fig. 9: Noticeable improvement of annotation quality through consecutive stages of ensemble training. The ground truth annotation is shown at the bottom, the 4th networks annotation is located in the middle and the 3rd networks annotation is at the top. The 3rd network mistakenly notes the P-wave, blue. The remaining components of the cycle (green and red colors) are defined by both networks correctly

Fig. 10 depicts how the ensemble itself fixes an error of the base network for an abnormal case (patient was taken from

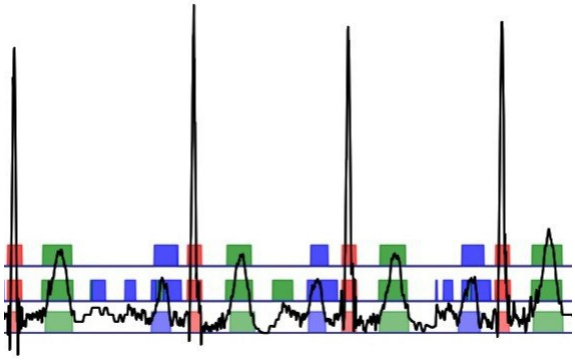


Fig. 10: Annotation for an abnormal case. The ground truth annotation is shown at the bottom, a single networks annotation is located in the middle and the ensembles annotation is at the top. The improved quality of segmentation since the introduction of the ensemble is clearly visible

test set). Table I shows that the ensemble of networks has improved most of the quality metrics compared to a single network. In next sections we will explore what caused these improvements.

A. *Overfitting analysis*

Networks added to the ensemble at later iterations were trained on very small patient subsets. This situation obliges us to check for the degree of overfitting in such networks.

In order to do so, we have designed a simple procedure, which roughly evaluates a degree of overfitting of all but one networks in an ensemble without need to use the test set. Every member of the ensemble (except for the first one) is only trained on some subset of the training set of patients. This allows us to use the unseen (for this member) part of the training set as a test set (for this member) and therefore to evaluate its generalization ability. Visualization of results is shown in fig. 11. It illustrates the behavior of an ensemble

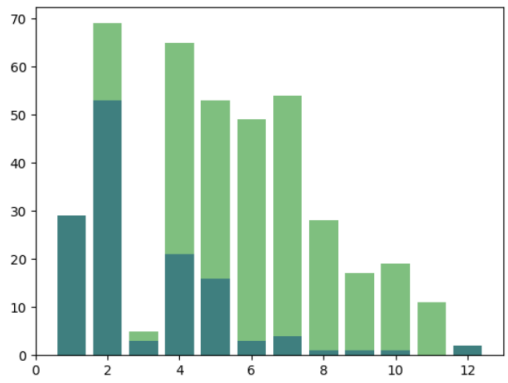
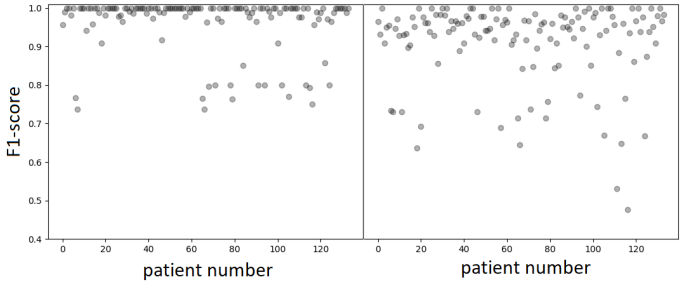


Fig. 11: Histogram of the generalization capability of every individual member of the ensemble. The ensemble members capability to annotate its part of the dataset exceptionally well (i.e. F1-score not less then 0.99) is shown in dark green, while its ability to annotate the previously unseen part of the dataset exceptionally well is shown in light green

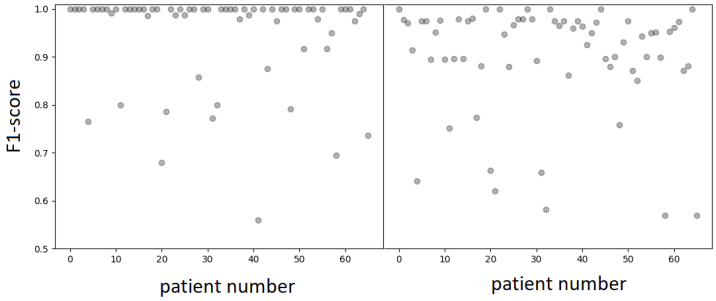
of 12 members. For every member of the ensemble the dark green bar shows the amount of well annotated patients from the training set of that member. At least half of members of the considered ensemble were trained on less than 10 patients (see fig. 8). The light green bar illustrates the amount of well annotated patients from the unseen part of the whole training set. We can see that there are only two networks which are probably significantly overfitted (dark green is too high comparing to light green). The first network does not have its own test set, so we cannot make any judgments about it. Some of the remaining ensemble members can produce good annotations (F1-score higher than 0.99) for tens of unseen patients despite having been trained on 2-3 patients. This is interesting in the light of the fact that deep learning models usually require large amounts of data for a good generalization.

B. *"Distillation" effect*

To get an idea of the behavior of an ensemble described above, we visualized its behavior on a test and training set. We then compared this behavior to the behavior of the single network. The resulting visualization is shown in Fig. 12.



(a) F1-score of segmentation given by an ensemble (left) or by single network (right) for the train set



(b) F1-score of segmentation given by an ensemble (left) or by single network (right) for the test set

Fig. 12: F1-score scattergrams for every patient demonstrates "distillation effect" looking from right to left. This behavior his behavior persists for both test set and train set.

The left subfigures demonstrate how the base network annotates testing and training sets. The right part depicts the same for the ensemble. The ensemble concentrates F1-scores

TABLE I: Quality metrics for the single network and ensemble. Values are averaged across 15 experiments

		P onset	P offset	QRS onset	QRS offset	T onset	T offset
single	Se(%)	95.20	95.39	99.51	99.50	97.95	97.56
	PPV(%)	82.66	82.59	98.17	97.96	94.81	94.96
	$m \pm \sigma$(ms)	2.7±21.9	−7.4±28.6	2.6±12.4	−1.7±14.1	8.4±28.2	−3.1±28.2
ensemble	Se(%)	97.97	97.36	99.86	99.95	93.77	93.51
	PPV(%)	90.48	90.72	98.27	98.66	96.28	96.23
	$m \pm \sigma$(ms)	3.4±18.4	−4.1±19.4	1.7±10.0	−3.4±12.3	9.2±28.2	−6.0±25.0

in a very narrow area near 1. I.e. after the ensemble processing the data set turned out to be divided into two classes: a dense cloud and a very rarefied one. The dense cloud consists of samples which were annotated by a single network with minor errors (F1-score near 0.9), and which after using the ensemble became annotated close to the ideal (F1-score near 1). We will call such examples "simple" and the remaining ones - "complex".

An example of a typical "complex" ECG is depicted in fig. 13.

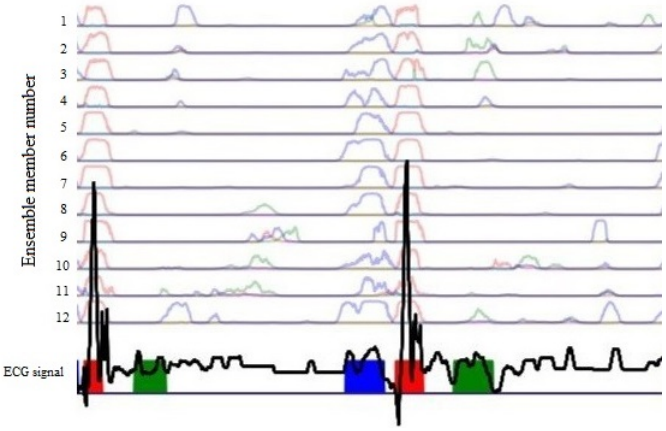


Fig. 13: An example of an outlier ECG on which both – the ensemble and the base network – fail: all the members have failed to segment the T-vawe (green), potentially because the amplitude of the T-wave has become comparable in scale to the amplitude of the noise. Colored blocks indicate the segmentation of the expert. The top markings show the output signals of the ensemble members

Overall F1-score of the ensemble is higher than that of the single network. For the single network it is 0.94 and for the ensemble is more then 0.95, but the proportion of outliers detected by the ensemble is not less (about 16,5% for test sets) than that for the single network. This means that the ensemble improved the quality metrics mostly not at the expense of "complex" cases.

IX. Conclusion and Results

In this paper, we presented a qualitative study of how the error correction of a single network by an ensemble occurs (using the example of a typical ensemble and an important applied task from medicine). The most important results are:

1) Error correction does not occur evenly across the entire data set, but two classes of cases are clearly distinguished - "simple" (easily corrected by the ensemble) and "complex" (practically not amenable to correction).
2) Ensemble error is minimal in healthy patients, and they belong to the above-described "simple" class
3) The presence of a rare class of disease in ECG in question doesn't necessarily result in the ECG sample falling into the class of "complex" samples
4) With appropriate selection of an ECG in a training sample it is possible to achieve that the network trained only for 2-3 patients builds a rather good generalization. This allows to take a fresh look at the situation with "complex" and "simple" samples: the fact that the example got into the "complex" class is not a predetermined by the rareness of a specific disease in a dataset

X. Discussion

The number of samples from the aforementioned class of "complex" cases can be used as a basis for assessing the quality of internal representation that a network of a given architecture can build. In the case considered, for example, it can be concluded that internal representation built by the single network has serious flaws despite the fact that formal quality metrics (such as specificity, positive predictive value and F1-measure) show relatively high values. Also a promising area of research could be the search for a reliable metric that assesses the quality of network representation while doesn't consider the network as a black box. The creation of such metrics probably requires the further development of mathematical theory for deep learning.

References

[1] A. Gorban, A. Golubkov, B. Grechuk, E. Mirkes, and I. Tyukin, "Correction of ai systems by linear discriminants: Probabilistic foundations," *Information Sciences*, vol. 466, pp. 303–322, 2018.

[2] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger, "Snapshot ensembles: Train 1, get m for free," *arXiv preprint arXiv:1704.00109*, 2017.

[3] Q.-s. Zhang and S.-C. Zhu, "Visual interpretability for deep learning: a survey," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 27–39, 2018.

[4] B. W. Yap, K. A. Rani, H. A. A. Rahman, S. Fong, Z. Khairudin, and N. N. Abdullah, "An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets," in *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*. Springer, 2014, pp. 13–22.

[5] A. Arnab, O. Miksik, and P. Torr, "On the robustness of semantic segmentation models to adversarial attacks," *arXiv preprint arXiv:1711.09856*, vol. 2, 2017.

[6] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," *arXiv preprint arXiv:1506.06579*, 2015.

[7] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[8] C. Eastwood and C. K. Williams, "A framework for the quantitative evaluation of disentangled representations," 2018.

[9] A. Kalyakulina, I. Yusipov, V. Moskalenko, A. Nikolskiy, A. Kozlov, K. Kosonogov, N. Zolotykh, and M. Ivanchenko, "Lu electrocardiography database: a new open-access validation tool for delineation algorithms," *eprint arXiv:1809.03393*, 2018.

[10] P. De Chazal, M. O'Dwyer, and R. B. Reilly, "Automatic classification of heartbeats using ecg morphology and heartbeat interval features," *IEEE transactions on biomedical engineering*, vol. 51, no. 7, pp. 1196–1206, 2004.

[11] G. Lenis, N. Pilia, A. Loewe, W. H. Schulze, and O. Dössel, "Comparison of baseline wander removal techniques considering the preservation of st changes in the ischemic ecg: a simulation study," *Computational and mathematical methods in medicine*, vol. 2017, 2017.

[12] P. Rajpurkar, A. Y. Hannun, M. Haghpanahi, C. Bourn, and A. Y. Ng, "Cardiologist-level arrhythmia detection with convolutional neural networks," *arXiv preprint arXiv:1707.01836*, 2017.

[13] Association for the Advancement of Medical Instrumentation, "NSI/AAMI EC57:1998/(R)2008 (Revision of AAMI ECAR:1987)," 1999.

[14] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.

[15] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.

[16] Y. Liu and X. Yao, "Ensemble learning via negative correlation," *Neural networks*, vol. 12, no. 10, pp. 1399–1404, 1999.