IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 3, 2019, accepted January 3, 2020, date of publication January 23, 2020, date of current version January 31, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2968900

Machine Learning and End-to-End Deep Learning for the Detection of Chronic Heart Failure From Heart Sounds

MARTIN GJORESKI[1,2], ANTON GRADIŠEK[1], BORUT BUDNA[1], MATJAŽ GAMS[1,2], AND GREGOR POGLAJEN[3,4]

[1]Jozef Stefan Institute, SI-1000 Ljubljana, Slovenia
[2]Jozef Stefan Postgraduate School, SI-1000 Ljubljana, Slovenia
[3]Advanced Heart Failure and Transplantation Program, Department of Cardiology, UMC Ljubljana, SI-1000 Ljubljana, Slovenia
[4]Medical Faculty, University of Ljubljana, SI-1000 Ljubljana, Slovenia

Corresponding author: Martin Gjoreski (martin.gjoreski@ijs.si)

This work was supported in part by the Slovenian Research Agency under Grant U2-AG-16/0672 - 0287 and by the EU project PlatformUptake.eu.

ABSTRACT Chronic heart failure (CHF) affects over 26 million of people worldwide, and its incidence is increasing by 2% annually. Despite the significant burden that CHF poses and despite the ubiquity of sensors in our lives, methods for automatically detecting CHF are surprisingly scarce, even in the research community. We present a method for CHF detection based on heart sounds. The method combines classic Machine-Learning (ML) and end-to-end Deep Learning (DL). The classic ML learns from expert features, and the DL learns from a spectro-temporal representation of the signal. The method was evaluated on recordings from 947 subjects from six publicly available datasets and one CHF dataset that was collected for this study. Using the same evaluation method as a recent PhysoNet challenge, the proposed method achieved a score of 89.3, which is 9.1 higher than the challenge's baseline method. The method's aggregated accuracy is 92.9% (error of 7.1%); while the experimental results are not directly comparable, this error rate is relatively close to the percentage of recordings labeled as ''unknown'' by experts (9.7%). Finally, we identified 15 expert features that are useful for building ML models to differentiate between CHF phases (i.e., in the decompensated phase during hospitalization and in the recompensated phase) with an accuracy of 93.2%. The proposed method shows promising results both for the distinction of recordings between healthy subjects and patients and for the detection of different CHF phases. This may lead to the easier identification of new CHF patients and the development of home-based CHF monitors for avoiding hospitalizations.

INDEX TERMS Chronic heart failure, deep learning, heart sounds, machine learning, PCG.

I. INTRODUCTION

Chronic heart failure (CHF) is a chronic, progressive condition underscored by the heart's inability to supply enough perfusion to target tissues and organs at the physiological filling pressures to meet their metabolic demands [1]. CHF has reached epidemic proportions in the population, as its incidence is increasing by 2% annually. In the developed world, CHF affects 1-2% of the total population and 10% of people older than 65 years. Currently, the diagnosis and treatment of CHF uses approximately 2% of the annual healthcare budget. In absolute terms, the USA spent approximately 35 billion USD to treat CHF in 2018 alone, and the costs are expected to double in the next 10 years [2]. Despite the progress in medical- and device-based treatment approaches in the last decades, the overall prognosis of CHF is still dismal, as 5-year survival rate of this population is only approximately 50%. In the typical clinical course of CHF, we observe alternating episodes of compensated phases, when the patient feels well and does not display symptoms and signs of fluid overload, and decompensated phases, when symptoms and signs of systemic fluid overload (such as breathlessness, orthopnea, peripheral edema, liver congestion, pulmonary edema) can easily be observed. During the latter episodes, patients often require hospital admission to receive treatment with intravenous medications (diuretics, inotropes) to achieve a

The associate editor coordinating the review of this manuscript and approving it for publication was Juan Liu.

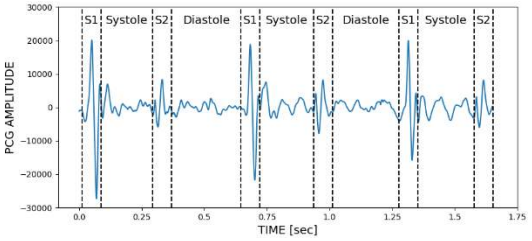

FIGURE 1. Example PCG recording with the PCG regions of interests (S1, S2, Systole and Diastole) marked.

successful negative fluid balance and return to the compensation state. Early detection of HF worsening would allow a treating physician to adjust the patient's medical management on an outpatient basis in a timely manner and thus avoid the need for a hospital admission. Currently, an experienced physician can detect the worsening of HF by examining the patient and by characteristic changes in the patient's heart failure biomarkers, which are determined from the patient's blood. Unfortunately, clinical worsening of a CHF patient likely means that we are already dealing with a fully developed CHF episode that will most likely require a hospital admission. Additionally, in some patients, characteristic changes in heart sounds can accompany heart failure worsening and can be heard using phonocardiography. An example of a phonocardiogram (PCG) recording of a healthy subject is presented in Fig. 1. In healthy subjects, 2 heart sounds are typically heard (called S1 and S2). S1 is caused by the closure of the mitral valve and ventricular wall in the early systole, S2 is caused by the closure of the aortic and pulmonary valves at the beginning of the diastole. Here, the interval between S1 and S2 is called systole, i.e., the contraction phase of the cardiac cycle, and the interval between S2 and S1 is called diastole, i.e., the relaxation phase of the cardiac cycle. Additional heart sounds (such as S3 and S4) can be heard in certain cardiac conditions and are never regarded as normal. In the case of CHF (in the course of decompensation), we can often hear a third sound (S3) that typically appears 0.1-0.2 s after the second sound, i.e., S2.

Recently, it has been demonstrated that some physiological parameters, such as the occurrence of additional heart sounds or increased blood pressure in the pulmonary circulation, already start to appear several weeks before the CHF patient develops a clinically evident decompensation episode. This is also an important therapeutic window where outpatient-based treatment interventions can reverse CHF deterioration and return the patient to the compensated state without the need for a hospital admission.

In recent years, many studies have proposed Machine-Learning (ML) approaches for the automatic detection of different heart conditions using PCG signals recorded with a digital stethoscope [1]. Nevertheless, methods that explicitly focus on CHF detection are quite scarce. The typical ML pipeline for the detection of different heart conditions is as follows: segmentation of the signals by detecting the ''typical'' heart sounds (i.e., S1 and S2), denoising of the signals, extracting individual frequency-domain and time-domain features, and learning a feature-based ML model (e.g., using ML algorithms, such as Random Forest or Support Vector Machine - SVM) that is capable of classifying healthy vs. unhealthy sounds. Most of the features currently used are based on medical and audio/signal analysis knowledge.

However, a PCG recording that sounds unhealthy to one expert may sound healthy to another one; therefore, doctors never diagnose a CHF patient using only heart sounds, but rather use a holistic view of the patient instead (i.e., extensive medical history, blood pressure, laboratory tests, etc.). This uncertainty is one reason why 9.7% of the recordings in the recent PhysioNet cardiology challenge [3] were actually labeled as ''unknown'' by experts, while the rest of the recordings were labeled as healthy or unhealthy.

The recent advancements in Deep Learning (DL) suggest that end-to-end learning (i.e., ML models that learn directly from the raw data and no features are needed) can outperform the classic, feature-based ML. For example, DL has achieved breakthrough performance in tasks such as pattern recognition problems [4], image processing [5], [6], natural language processing [7], [8], speech and audio processing [9], [10], and sensor data processing [11], [12]. For CHF detection, a successful combination of classic ML and end-to-end DL can outperform each single approach [13]. The classic ML approach learns from a large body of expert-defined features, and the DL approach learns both from a time-domain (the raw PCG signal) representation of the signal and a temporal-domain representation (the spectrogram) of the signal. This approach was successful in our previous study of human activity recognition from smartphone sensor data [14].

In addition to distinguishing the CHF patients and healthy individuals, we focus on detecting the CHF state (compensated vs. decompensated) based on the analysis of heart sound recordings. Our work builds upon the initial studies, where we demonstrated that it is possible to distinguish between healthy individuals and patients in a decompensated CHF episode using a stack of machine-learning classifiers and expert features, showing promising results on a limited dataset [15]. We expand upon this approach using a considerably larger patient dataset, including six additional PhysioNet datasets, and an improved ML method that uses end-to-end DL. Furthermore, we investigate the differences in the heart sounds during the transition between the decompensated and recompensated states of CHF, with the aim of developing personalized monitoring models. Early detection of the worsening of CHF has the potential to reduce hospitalizations due to the worsening of the condition, which both improves the quality of life of patients and decreases the financial and logistic burden on the patient and the health system.

II. RELATED WORK

A typical ML pipeline includes segmentation of the signals, denoising of the signals, extracting individual frequency-domain and time-domain features, and

learning ML capable of classifying healthy vs. unhealthy sounds [1], [3]. Regarding the segmentation process, Schmidt *et al.* [16] developed an algorithm (later improved by Springer *et al.* [17]) that segments the signals into the following four stages: S1, S2, systole, and diastole. The algorithm extracts a variety of features that are then used to train a duration-dependent hidden semi-Markov model to segment the PCG. To ease the segmentation process, some researchers apply denoising techniques to remove environmental sounds and the noises caused by the human body itself. The next phase in the ML pipeline is the feature extraction, as the features are the basis for a successful classification. Most researchers focus mainly on time, frequency, and statistical features. The widely used features are as follows: heart rate, duration of S1, S2, SYS or DIA, total power of the PCG signal, zero crossing-rate, Mel-frequency Cepstral Coefficients, Wavelet Transform, Linear Predictive Coefficients, and Shannon entropy [18], [19].

The final phase in the ML pipeline is learning and evaluation of the ML models. The most systematic comparison of the ML models was performed via the PhysioNet challenge [1]. The challenge aimed to encourage the development of algorithms to classify heart sound recordings collected from a variety of clinical or nonclinical (such as in-home visits) environments. More details about the challenge dataset can be found in section *2.2* of the *PhysioNet datasets*. During the challenge, the ML models were ranked using an average of the weighted-sensitivity and the weighted-specificity scores achieved by the models. The weights were used as a normalization factor for the noisy recordings in the data. The best score of 86.0 was achieved by Potes et al. [20] using a method that was based on an ensemble of classifiers combining the outputs of an AdaBoost classifier and a Convolutional Neural Network (CNN). The second-best result, which was 85.9, was achieved by Zabihi *et al.* [21] using an ensemble of feature-based feedforward neural networks. Similarly, Kay and Agarwal [22] used a fully connected, neural network and achieved a score of 85.2. In fourth place, Bobillo [23] achieved a score of 84.5 using a tensor technique. Homsi *et al.* [18], who achieved a score of 84.5, introduced an approach using a nested ensemble of algorithms that includes Random Forest, LogitBoost and a Cost-Sensitive Classifier. A probabilistic approach based on logical rules and a probability assessment was proposed by Plesinger et al. [19], which achieved a score of 84.1. Rubin *et al.* [24], who achieved a score of 84.0, used CNNs on MFCC heat maps.

Although the challenge datasets present a great opportunity to compare methods for classifying heart sounds, unfortunately, the challenge did not specifically include recordings from CHF patients, and second, it does not provide full access to the challenge test datasets. Thus, we cannot make any CHF-detection comparison. However, we used the publicly available challenge datasets[1] for evaluation and compared the results to the challenge baseline method [3].

[1]https://www.physionet.org/content/challenge-2016/1.0.0/

With respect to the related work, our approach differs in the following aspects. (i) Most of the approaches from the PhysioNet challenge used the algorithm developed by Schmid *et al.* [16] for the initial segmentation of the PCG signals, which is based on the detection of the typical heart sounds. However, in noisy environments, the detection of these sounds may be even more challenging than the classification itself. For that reason, our method does not use such a strict segmentation, but rather an overlapping-sliding window technique in combination with a *segment-based* classifier. We present the analysis of the method's performance with respect to the segment (window) size. (ii) The proposed end-to-end DL architecture learns both from the temporal representation of the signal and the spectral representation of the signal, whereas most of the approaches in the related works use end-to-end learning in one of the domains only (either spectral or temporal [14]). (iii) We used the PhysioNet Challenge datasets to evaluate our approach and to provide a comparison with the challenge baseline method; additionally, we used our own dataset, which, in addition to including the typical healthy vs. patient labels, is also labeled for the specific CHF phase, i.e., compensated (when the patient feels well) and decompensated (when the patient does not feel well) for some of the patients. This allowed us to extend the study beyond the typical healthy vs. patient analysis and to explore personalized models for detecting the different CHF phases. To the best of our knowledge, this is the first computer-science study developing CHF detection models.

III. DATA DESCRIPTION

A. UKC-JSI DATASET

Seven datasets were used in our study. The first was collected by the authors of this paper, while the rest of the datasets originated from the 2016's PhysioNet challenge [1]. Our dataset (UKC-JSI, Table 1) was obtained using a professional digital stethoscope 3M^{TM} Littmann Electronic Stethoscope Model 3200. The stethoscope allows the recording of up to 12 clips of up to 30 s in length, with a sampling rate of 4 KHz. The device uses built-in filters to reduce the ambient noise and allows different settings to focus on listening to heart or lungs – for our experiments, we always opted for the option with the minimal filtering. The recordings were transferred to a computer via a Bluetooth connection and were analyzed offline.

The study was approved by the medical ethics committee beforehand. We recorded 110 healthy people (meaning that they had no medical condition) and 51 people diagnosed with CHF. For 22 CHF patients, recordings were obtained both during the decompensation episode (when hospitalized) and during the compensated phase (when discharged).

The recordings were always obtained at Erb's point, and each recording was up to 30 s long (stethoscope's limit). For some healthy people, more than one recording was obtained to increase the amount of data in the study (recordings of patients were obtained in clinical settings, which limited the available time).

TABLE 1. Overview of our experimental data recorded on healthy individuals and on patients in decompensated and recompensated CHF episodes.

	Decomp.	Recomp.	Healthy	All
# Subjects	51	22	110	183
# Recordings	52	22	159	233
Duration (min)	17	7	52	76

TABLE 2. Overview of the six (A to F) PhysioNet Challenge datasets.

			Proportion (%)		
Data	#Subjects	#Recordings	Abnorm.	Normal	Unsure
A	121	409	67.5	28.4	4.2
B	106	490	14.9	60.2	24.9
C	31	31	64.5	22.6	12.9
D	38	55	47.3	47.3	5.5
E	356	2054	7.1	86.7	6.2
F	112	114	27.2	68.4	4.4
ALL	764	3153	18.1	73.0	9.7

B. PhysioNet DATASET

The PhysioNet 2016 Cardiology Challenge database consists of six datasets (A through F, recorded by six research groups - participants of the competition) containing a total of 3,153 heart sound recordings, lasting from 5 seconds to just over 2 minutes. The recordings were obtained in either a clinical or a nonclinical environment, from both healthy subjects and pathological patients, and from different locations on the body. The four typical locations were the aortic area, the pulmonic area, the tricuspid area, and the mitral area, but could be one of nine different locations. Both the healthy subject and pathological patient samples include both children and adults. Each subject/patient may have contributed between one and six recordings. However, there is no information in the dataset regarding which recording belongs to whom. All recordings were resampled to 2,000 Hz and stored in the ''wav'' format. Each recording contains only one PCG lead. As the recordings were often collected in uncontrolled environments, many of the recordings are corrupted by various noise sources, such as talking, stethoscope motion, breathing, and intestinal sounds. In addition, some recordings were difficult or even impossible to classify as normal or abnormal. The summary of the recordings in the Challenge database can be found in Table 2. In addition, each of the six teams recording the training sets also produced a testing set, which was used for the evaluations of the contributed algorithms. We did not have access to the testing sets; therefore, we do not include them in our analysis.

IV. METHOD

The outline of our method is presented in Fig. 2. It consists of the following two main components: a classic ML component (represented with colored squares on the right side of the

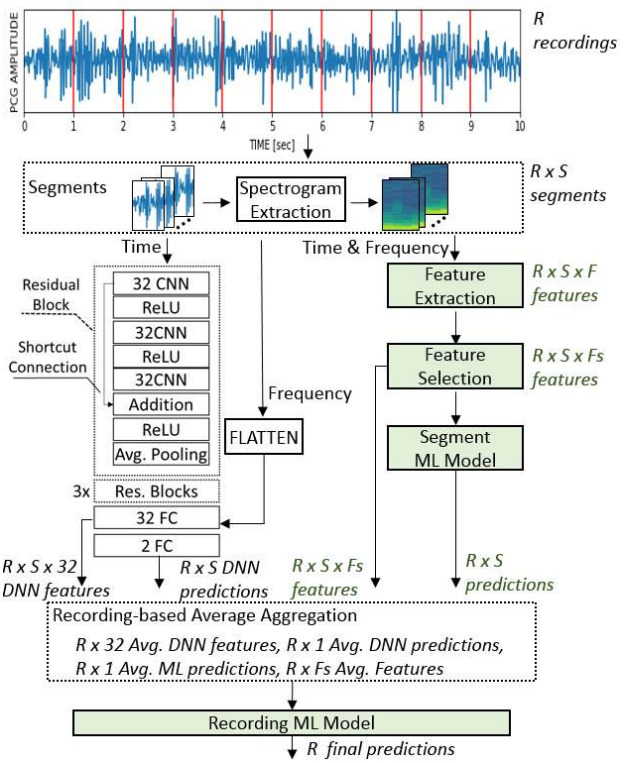

FIGURE 2. Proposed method. End-to-end DL (uncolored squares on the left). Classic ML (colored squares).

figure) and an end-to-end DL component (represented with noncolored squares). The input to the classic ML pipeline is the same as the input to the end-to-end DL pipeline, but the classic ML pipeline contains a feature extraction process to extract features from the raw data and to format the data into a classic ML format. The end-to-end DL does not require feature extraction, and it works directly with the raw data. Additionally, both pipelines work with signals from the time and frequency domains. The outputs of the two components are then merged by a *recording-based* ML model that outputs the final prediction, i.e., whether a recording comes from a healthy subject or from a patient. The details of the method are explained in the following subsections.

A. CLASSIC MACHINE LEARNING

The classic ML component consists of feature extraction, feature selection, and a *segment-based* ML model. As noises not related to heart sounds are expected to be present in the recordings, the first step is filtering. Most cardiovascular sounds are most likely to occur in the frequency range below 1 kHz [25]; thus, we applied a low-pass Butterworth filter with a threshold of 1 kHz to the raw audio files. For the segmentation of the filtered audio signal, we used a sliding window technique.

Next, the audio features were extracted using the OpenSMILE feature extraction tool [26]. The tool was originally created for acoustic emotion recognition in 2009 but was later expanded to more general uses [27]. For example,

in addition to affect recognition, it is widely used for music information retrieval (e.g., chord labeling, beat tracking, etc.).

Based on the related work [18], [20], [21], we extracted features from the recordings in the time and frequency domains. For example, Potes *et al.* [20] extracted features in the time and frequency domains. The time-domain features include the mean, variance, skewness, kurtosis, etc. The frequency domain features include the power across different frequency bands, 13 MFCCs, etc. Similarly, Zabihi *et al.* [21] used features in both the time and frequency domains. Their features include the following: Linear Predictive Coefficient (LPC), Entropy-based features, MFCCs, features extracted over power spectral density, etc. In addition to the typical time- and frequency-based features, Homsi et al. [18] added entropy and the Zero Crossing Rate (ZCR). Additionally, we included features that describe the noise in the signal, e.g., jitter, shimmer and the Harmonics-to-Noise Ratio.

The complete list of features is described by Florian and Schuller et al. [26], [28], [29]. For each segment, 1941 features were extracted on a segment level and 1941 additional features were extracted on a recording-level. Thus, each segment is represented by segment-based features, which describe only the segment itself, and recording-based features, which describe the whole recording. The shortest segment size used in our experiments was one second so that each segment contained at least one complete heartbeat (with common heart rates above 60 beats per minute). By extracting features both form the segments and from the whole recordings, we aim to capture both short-term information and the long-term information. The short-term information from the segment-based features mainly describes the heart beats, and the long-term information describes the overall recording. The output of the feature extraction component is marked as "$R \times S \times F$ *features*" in Fig. 2, since we obtain F-features for each R-recording and S-segment.

As the number of available features for each segment (1941×2) is several times larger than the number of available recordings in most of the experimental datasets, a feature selection step is required to avoid overfitting. In general, the feature selection methods can be divided into wrapper methods, ranking methods (also known as filter methods) and a combination of the two. The wrapper methods (e.g., based on ROC [30], Bhattacharyya distance [31], etc.) produce better results compared to the ranking methods (e.g., methods based on mutual information [32], information entropy [33], Fisher score [34], Wilcoxon signed-rank test [35], etc.), but they induce a heavy computation burden. We decided to use mutual information because it is very efficient and fast to compute. However, any other feature selection or dimensionality reduction methods (e.g., PCA [36]) can be used.

Mutual information is a measure that estimates the dependency between two random variables. We ranked the features using mutual information values between the features and the class values. We used only the top-ranked m features, where m was set to be equal to the number of training samples. The features were ranked using 10% of the instances from the training data, which were randomly selected. The output of the feature selection component is marked as "$R \times S \times Fs$ *features*" in Fig. 2.

After the feature selection, a *segment-based* classifier was trained. The *segment-based* classifier uses the segments as input instances, represented via the selected features from the previous step, and outputs the estimated class probabilities for each segment (segment-class probabilities) of each recording. The output probabilities are marked as "$R \times S$ *predictions*" in Fig. 2. The segment-class probabilities are later used as the input to the *recording-based* classifier. The *segment-based* classifier was trained using the Random Forest (RF) algorithm. We chose the RF algorithm because it is robust to noise in the input features. The *recording-based* classifier is described in the *Combining Classic ML and end-to end DL* section.

B. DEEP LEARNING

DL represents a class of ML algorithms that use a cascade of multiple layers of nonlinear processing units [37]. The first layer receives the input data, and each successive layer uses the output from the previous layer as input. DL architectures are able to solve complicated AI tasks (e.g., in computer vision, language, biomedicine, etc.) by learning high-level abstractions from raw data [38].

A cascade of multiple layers of nonlinear processing units, where each processing unit receives input from the previous layer, is called a Fully Connected Neural Network (FCNN). In a typical FCNN, layer i computes an output vector z_i using the following equation:

$$\mathbf{z_i} = f(\boldsymbol{b_i} + \boldsymbol{W_i}\mathbf{z_{i-1}}) \tag{1}$$

where b_{i} (biases) and W_{i} (weights) are the parameters for the i^{th} layer. z_{i-1} is the output vector of the previous layer, and z_0 is the input data. The activation function f can be a Rectified Linear Unit (ReLU), as follows [39]:

$$f\ (\mathrm{c}) = \max\ (0, \mathrm{c}) \tag{2}$$

or some other nonlinear function, for example, sigmoid or tanh. For classification problems, the final output layer (z_i) usually uses a softmax activation function, as follows:

$$\mathrm{z}_{i_j} = \frac{e^{b_{i_j} + W_{i_j}\mathrm{z}_{i-1}}}{\sum_j e^{b_{i_j} + W_{i_j}\mathrm{z}_{i-1}}} \tag{3}$$

where j represents the j^{th} row of the weights W_i. The softmax function has a nice property, as follows:

$$\sum\nolimits_j z_{I_j} = 1 \tag{4}$$

and it is always positive; thus, it can be used as an estimator for an input data sample x to belong to the j^{th} class for a specific problem, as follows:

$$P(y = j|\boldsymbol{x}) \tag{5}$$

The parameters of the network (b_i and W_i) are learned using an optimization algorithm, for example, gradient

descent [40]. For a binary classification problem (e.g., y is either 1 or 0), binary cross-entropy is used as a loss function that is minimized over the N pairs of data samples and labels (x_n, y_n).

$$J(\boldsymbol{W}) = -\frac{1}{N}\sum\nolimits_n [y_n log\,(p_n) + (1 - y_n)\,log(1 - p_n)] \quad (6)$$

where, p_n is the estimated probability for the N^{th} sample to belong to class 1.

CNNs are a type of NN that are designed with three main architectural ideas to ensure some degree of shift, scale, and distortion invariance. This is achieved by utilizing the following: (i) local receptive fields, i.e., each unit in a layer receives input from a set of neighboring units in the previous layers; (ii) shared weights (units in a layer are organized in groups and all units in one group share the same set of weights); the set of outputs of the units in one group is called a feature map, and the set of connection weights used by the units to create the feature map is called a kernel or filter [41], [42]; (iii) spatial or temporal sampling, where, if the input is shifted, the feature map output will also be shifted [43]. In addition, because of the specific architecture (parameter sharing and local connections), the CNNs have much fewer connections and parameters to train, while their theoretical-best performance is likely to be only slightly worse compared to that of FCNNs [5].

The DNN architecture used in this study is deep *Spectro-temporal ResNet*. A similar *Spectro-temporal ResNet* architecture has already proved successful for human activity recognition in our previous study [14] by achieving comparable accuracy to state-of-the-art feature-based models. The structure is based on an idea for training very deep end-to-end networks for image recognition; i.e., it uses shortcut (residual) connections to fight the gradient-vanishing problem [44]. Additionally, our architecture consists of two branches, i.e., one that works with the raw PCG signal in the time domain and another that works with the spectral representation of the signal. The temporal information is extracted by residual blocks that contain 1D CNN filters. To reduce the internal covariance shift, each CNN layer is followed by a batch normalization layer [45]. To speed up the training process, ReLU activation layers are used [46]. For the dimensionality reduction, each residual block ends up with a maximum pooling layer. The network obtains the spectral information by calculating a spectrogram of the input signal. Toward the end of the network, the two branches, i.e., the spectral and the temporal branch, are merged using FC layers. The output of the end-to-end DL component is marked as "$R \times S$ *DNN predictions*", as it outputs one prediction for each segment (S) of each recording ®. Additionally, the output of the second to last FC layer is marked as "$R \times S \times 32$ *DNN features*", as it contains 32 hidden units; thus, it outputs a vector with a size of 32 for each segment. This vector represents the spectro-temporal encoding of each segment and is also the input to the *recording-based* ML model.

The FC layer learns a spectro-temporal encoding, which is further utilized by the *recording-based* ML model. We used two FC layers, each with a size of 32 units; thus, the spectro-temporal encoding is represented by a vector of a size of 32.

The *Spectro-Temporal ResNet* was trained by minimizing the binary cross-entropy loss function using the Adam optimizer with a learning rate of 10^{-3} and a decay of 10^{-3}. The batch size was set to 256, and the maximum number of training epochs was set to 20. The network parameters, including the number of residual blocks, the number of CNN layers per block, the size of the CNN filters, the learning rate, and the batch size, were determined experimentally.

C. COMBINING CLASSIC MACHINE-LEARNING AND END-TO-END DEEP LEARNING

The four outputs of the components before the recording-based ML model ("$R \times S \times Fs$ *features*", "$R \times S$ *predictions*", "$R \times S \times 32$ *DNN features*" and "$R \times S$ *DNN predictions*") are first averaged for each recording (thus, we obtain the averaged "$R \times Fs$ *features*", "R *predictions*", "$R \times 32$ *DNN features*" and "*R DNN predictions*") and are then used as the input to the recording-based ML model. Finally, the *recording-based* ML model outputs the final prediction for each recording. The motivation here is the fact that all the segments in a chosen recording belong to the same class. However, some segments may be more informative than others; therefore, a segment-based classification followed by the aggregation of segment-based forecasts should improve the overall classification. The *recording-based* classifier was trained using the RF algorithm.

Each of the three classifiers (the *segment-based* classifier, the *Spectro-temporal ResNet* and the *recording-based* classifier) were trained separately on the training data. Since the *recording-based* classifier is a meta-learner that utilizes the output of the *segment-based* classifier and the *Spectro-temporal ResNet*, a holdout set is required for its training. More specifically, we used a 10-fold cross-validation process on the method's training data to train the ML and DL segments. During the 10-fold cross-validation process, both the *segment-based* classifier and the *Spectro-temporal ResNet* provided output for each of the 10 folds. The recording ML was trained on those outputs. Note that once trained, the whole method was evaluated on a separate test set, which was not used in the training phase of the three ML models.

D. BASELINE METHOD

The baseline method starts with heart sound segmentation using the Springer's algorithm [17] for detecting the four states, i.e., S1, systole, S2, and diastole (see Fig. 1). After that, it extracts twenty features from the position information of the four states mentioned above. The first ten features are defined as the averages and standard deviations of the beat-to-beat intervals (RR intervals), S1, S2, systolic and diastolic intervals. The last ten features describe the averages and standard deviations of the ratios between the different intervals and the ratios between the mean absolute amplitude

during systole or diastole to that during the S1 or S2 period in each heartbeat. The features are fed to a binary logistic regression classifier. To provide a better comparison, we also used an ensemble method (Random Forest) as another baseline classifier.

V. EXPERIMENTS

We performed three types of experiments. In the first set of experiments, i.e., the *PhysioNet experiments*, we evaluated the method for classifying healthy vs. patient recordings on seven datasets and compared the results against the baseline methods. In the second set of experiments, i.e., the *UKC-JSI experiments*, we evaluated the method for different window sizes and performed subject-independent evaluation for the classification of healthy vs. patient recordings. In the third set of experiments, i.e., the personalization experiments, we analyzed the expert-features for the classification of recompensated vs. decompensated recordings on the UKC-JSI dataset and performed a subject-independent evaluation of a model built with those features. In these experiments, we used only the expert-features, as the number of recompensated/ decompensated recordings is fairly small (44) for training an end-to-end DL model. For the evaluation of the ML models, we used the following evaluation metrics (in percent):

$$Accuracy = 100 * \frac{Correctly\ classified\ recordings}{Number\ Recordings} \quad (7)$$

$$Sensitivity = 100 * \frac{Correctly\ classified\ Patient\ recordings}{Patient\ recordings} \quad (8)$$

$$Specificity = 100 * \frac{Correctly\ classified\ Healthy\ recordings}{Healthy\ recordings} \quad (9)$$

$$Score = Average(Sensitivity, Specificity) \quad (10)$$

A. PhysioNet EXPERIMENTS

In these experiments, we used the same evaluation approach that the organizers of the PhysioNet Challenge used to evaluate their baseline model [3]. The evaluation procedure is a recording-independent 10-fold cross-validation process, which means that one recording can be either only in the training dataset or only in the testing dataset. We performed the 10-fold evaluation for each dataset specifically. The folds were not randomized to ensure that each recording is used as a test recording at least once. The 10-fold evaluation was performed 10 times for each dataset, and the average accuracy for each dataset is presented in Table 3. For each of these datasets, our method – the *recording-based* classifier (*Rec.* in Table 3) – outperformed the other classifiers (the majority classifier, the *baseline LogReg* classifier, the *baseline RF* classifier and the *segment-based* classifier (*Seg.* in Table 3). For the *segment-based* classifier, the prediction for a recording is calculated as a majority of the predictions of its segments. Additionally, the accuracy highly depends on the dataset.

TABLE 3. Accuracy for each of the PhysioNet Challenge datasets (A-F) and for our dataset UKC-JSI (U).

	A	B	C	D	E	F	U
Majority	71.4	78.8	77.4	50.9	92.4	70.2	66.6
LogReg	72.4	79.6	77.4	67.3	92.6	72.8	64.1
RF	69.4	77.6	87.1	72.7	92.9	74.6	74.2
Seg.	77.6	80.9	90.3	89	**99.7**	72.7	76.8
Rec.	**80.4**	**82.5**	**100**	**92.6**	**99.6**	**77.6**	**84.6**

TABLE 4. p-values from the McNemar's tests between our method (Segment-based) and the other methods.

	Segment-based						
	A	B	C	D	E	F	U
LogReg	0.028	0.000	0.000	0.002	0.038	0.053	0.000
RF	0.024	0.004	0.004	0.031	0.043	0.167	0.004
Seg.	0.043	0.265	0.045	0.375	0.465	0.013	0.005

More specifically, the *recording-based* classifier achieved the lowest accuracy, i.e., 77.6%, for dataset F and the highest accuracy, i.e., 99.6%, for dataset E. For our dataset, the hboxUKC-JSI dataset (U in Table 3), the accuracy is 84.6%.

It is interesting to note that, for the datasets C, D, and F, the number of participants is almost equal to the number of recordings (see Table 2); thus, the evaluation is almost the same as a person-independent 10-fold cross-validation process (one person can be either only in the training dataset or only in the testing dataset). This means that for these datasets, the models are roughly person-independent.

The results presented in Table 3 show that our method achieves the highest accuracy, which indicates the practical significance of the results. To check for statistical significance, we used the McNemar's statistical test, which is a recommended statistical test for comparing two classifiers over one dataset because it has a low type I error [47], [48]. For each dataset, we performed pairwise comparisons between our method and each of the other methods. More specifically, the test compares the prediction errors made by both models and checks whether there is a significant difference between them. The p-values of these tests are presented in Table 4. It can be observed that the differences between our method and the *baseline RF* and *baseline LogReg* are statistically significant ($p < 0.05$) for six out of the seven datasets. When compared to the recording-based method, the differences are significant for four out of the seven datasets, which is expected because the *recording-based* classifier is one module of our method (the *recording-based* method). Most importantly, for the UKC-JSI dataset (U in Table 4), there are statistically significant differences between our method and all of the comparison methods (with p-values of 0.005, 0.004 and 0.000).

Table 5 presents a normalized confusion matrix, i.e., the rows sum up to 100, as well as the sensitivity, specificity, and score for the UKC-JSI dataset using the *recording-based* classifier and a recording-independent 10-fold cross-validation process.

TABLE 5. Normalized confusion matrix (first two rows and columns), sensitivity, specificity, and score (average value of the sensitivity and specificity) for the UKC-JSI dataset using the recording-based classifier recording-independent 10-fold cross-validation.

	Healthy	Patient	Spec.	Sens.	Score	Acc
Healthy	93.5	6.5				
Patient	33.7	66.3	93.5	66.3	79.9	84.2

TABLE 6. Aggregated confusion matrix (first two rows and columns) over all datasets; sensitivity, specificity and score (average value of the sensitivity and specificity).

		Healthy (H)	Patient (P)	Spec.	Sens.	Score	Acc
LogReg	H	93.1	6.9	93.1	50.1	72.0	83.1
	P	49.9	50.1				
RF	H	93.6	6.4	93.6	66.8	80.2	86.7
	P	33.2	66.8				
Seg.	H	56.7	43.3	56.7	73.7	65.2	60.8
	P	26.3	73.7				
Rec.	H	96.2	3.8	**96.2**	**82.3**	**89.3**	**92.9**
	P	1	82.3				

TABLE 7. Accuracy for the recording-based classifier for varying window sizes on the UKC-JSI dataset.

Window size in seconds					
10	8	6	4	2	1
82.6	81.6	82.9	84.9	**86.1**	84.6

The results suggest that the sensitivity is lower than the specificity, which means that the classifier produces more false negatives in these experiments.

To present a clearer comparison with the scoring method from the PhysioNet Challenge, in Table 6 we present the aggregated (and normalized) confusion matrix over all the datasets for each classifier, and the achieved sensitivity, specificity, and score. Again, it can be observed that the *recording-based* classifier is the best performing classifier, with a score of 89.3 and an accuracy of 92.9%.

B. UKC-JSI EXPERIMENTS

In these experiments, we analyzed the relation between the window size and the performance of the *recording-based* classifier. For that reason, the experiments were performed using varying window sizes (10, 8, . . . , 2, and 1 second) and a 50% overlap. One second was chosen as the shortest window so that each segment contains at least one complete heartbeat (with common heart rates above 60 beats per minute). The evaluation was performed using the same recording-independent 10-fold cross-validation process as before to obtain comparable results. We only used the UKC-JSI dataset for these experiments, as our goal is to tune the algorithm for this dataset. The results are presented in Table 7.

Finally, the best performing classifier – the *recording-based* classifier using a sliding window of 2 seconds and a 50% overlap – was evaluated using a Leave-One-Subject-Out (LOSO) evaluation procedure. This evaluation produces more a reliable estimate of the method's performance for a new (unseen) subject. The LOSO evaluation is a type of person-independent k-fold cross-validation, where k is equal to the number of participants in the dataset. We could carry out the LOSO experiment only on the UKC-JSI dataset since the PhysioNet dataset does not specify which recording belongs to which individual. The results are presented in Table 8. The accuracy of 86.3% is similar to the accuracy from the previous experiments (84.6%), which confirms that overfitting was avoided in the previous evaluation. Similarly, as in previous experiments, the sensitivity is lower than the specificity.

TABLE 8. Normalized confusion matrix (first two rows and columns), sensitivity, specificity, and score (average values of the sensitivity and specificity) for the UKC-JSI dataset using the recording-based classifier and a leave-one-subject-out evaluation.

	Healthy	Patient	Spec.	Sens.	Score	Acc
Healthy	94.6	3.4				
Patient	28.5	71.5	94.6	71.5	83.0	86.3

C. PERSONALIZATION

For 22 out of 51 patients in the UKC-JSI dataset, we obtained one recording in the decompensated phase, i.e., when the patient needs medical attention, namely, at the time of hospital admission, and one in the recompensated phase, i.e., when the patient feels well and is released from the hospital. We analyzed whether we can distinguish between the two phases, as this is the first step toward the long-term goal of building a model that would allow us to continuously monitor the worsening of the CHF condition.

We preformed statistical tests to check whether there is a difference in the features when calculated from the recompensated recordings compared to the decompensated recordings. We used the Wilcoxon signed-rank test, a nonparametric statistical hypothesis test used to determine whether two paired samples are sampled from the same distribution [49]. In this experimental setting, one sample contains the values of a specific feature extracted from the recompensated recordings and the other sample contains the values for the same feature but extracted from the decompensated recordings. The informative features should have different distributions depending on the type of recording. Table 9 presents the 15 features for which the tests showed p-values smaller than 0.001.

Finally, we built a transparent ML model that can distinguish between the two phases. Since the dataset, which contains 44 recordings from 22 patients, is too small, we used a simple decision tree classifier conditioned to a maximum depth of two, to minimize overfitting. For training the model, we used the features in Table 9.

Before feeding them to the algorithm, the features were normalized by subtracting the person-specific mean value from each feature. Table 10 presents the evaluation results (normalized confusion matrix, sensitivity, specificity, score, and accuracy). Only 3 out of 44 recordings were

TABLE 9. features that showed a statistically significant difference ($p<0.001$) in the distribution depending on the CHF phase from which they were extracted.

Name (in OpenSmile [28])	Explanation
pcm_fftMag_spectralRollOff90.0**_sma_stddev** pcm_fftMag_spectralRollOff90.0**_sma_de_percentile1.0** pcm_fftMag_spectralRollOff90.0**_sma_de_percentile99.0** pcm_fftMag_spectralRollOff90.0**_sma_de_pctlrange0-1** pcm_fftMag_spectralRollOff90.0**_sma_de_posamean** pcm_fftMag_spectralRollOff90.0**_sma_de_rqmean** pcm_fftMag_spectralRollOff90.0**_sma_de_stddev**	**Segment based.** Statistical descriptors of the **spectral roll-off**, which is a measure of the amount of the right-skewedness of the power spectrum. "sma" represents moving average filter. "de" represents the first differential of the values. The specific statistical descriptors are as follows: standard deviation, 1st and 99th percentile, percentile range, positive arithmetic mean, root quadratic mean and standard deviation).
pcm_fftMag_spectralRollOff90.0**_sma_de_iqr1-3** pcm_fftMag_spectralRollOff90.0**_sma_de_percentile99.0** pcm_fftMag_spectralRollOff90.0**_sma_de_flatness** pcm_fftMag_spectralRollOff90.0**_sma_de_posamean** pcm_fftMag_spectralRollOff90.0**_sma_de_rqmean**	**Recording based.** Statistical descriptors of the spectral roll-off. The specific statistical descriptors are as follows: interquartile range, 99th percentile, flatness, positive arithmetic mean and root quadratic mean.
pcm_fftMag_psySharpness_sma_**de_percentile99.0** pcm_fftMag_psySharpness_sma_**de_pctlrange0-1**	**Segment based.** 99th percentile and percentile range of the psychoacoustic sharpness. A feature that quantifies some spectral characteristics. A high-frequency signal has a high value of sharpness
pcm_fftMag_spectralEntropy_**sma_percentile99.0**	**Segment based.** 99th percentile of the spectral entropy.

TABLE 10. Confusion Matrix for classifying Decompensated vs. Recompensated using a simple Decision Tree classifier and a leave-one-subject-out evaluation.

	Patient	Healthy	Spec.	Sens.	Score	Acc
Healthy	95.5	4.5				
Patient	9.1	90.9	95.5	90.9	93.2	93.2

misclassified, which lead to sensitivity, specificity, score, and accuracy values higher than 90%.

The decision tree model built in the final iteration of the LOSO evaluation is presented in Fig. 3. From the figure, we can see that the model is quite simple. By using just two features (X3 - spectralRollOff90.0_sma_de_pctlrange 0-1 and X9- pcm_fftMag_spectralRollOff90.0_sma_de_ flatness), the training data can be divided in a way where only one sample is misclassified. More specifically, at the beginning, there are 42 training samples (21 subjects, each with 2 samples) and 2 test samples. The "gini" value is a measure of how often a randomly chosen sample would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset of the specific node. By checking the value of the feature X_3, the model splits the samples into two subsets, where one subset has a feature value smaller than the specific value (-0.345) and the other subset

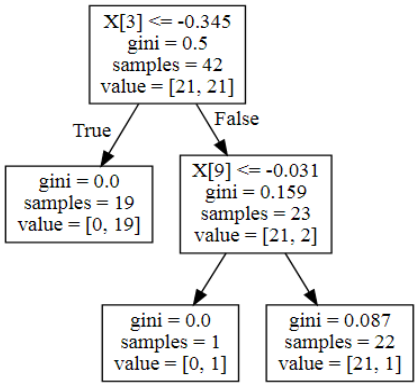

FIGURE 3. Decision tree for classifying the decompensated vs. recompensated phases, built in the final iteration of the LOSO evaluation.

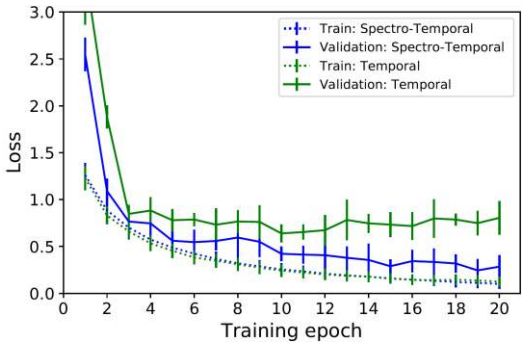

FIGURE 4. Averaged learning curves (loss values) and standard deviations for the end-to-end DL pipeline with spectrograms (Spectro-Temporal blue) and without spectrograms (Temporal green). 10-fold cross validation.

has a feature value larger than the specific value. The samples that have a smaller feature value (19 out of 42 samples) are classified as "1" (recompensated recordings). The samples that have a larger value are further split by using one more feature, i.e., X_9. After the second split, again, the samples that have a smaller feature value are classified as "1", and the rest of the samples are classified as "0" (decompensated).

VI. DISCUSSION

To analyze the influence of the spectrograms in the end-to-end DL component, we analyzed the learning curves (average loss values and standard deviations) for the end-to-end DL pipeline with spectrograms (Spectro-Temporal DL) and those without spectrograms (Temporal DL). The results are presented in Fig. 4. It can be observed that both models achieved low loss values with the training data. However, for the testing data, the Spectro-Temporal model achieves significantly lower loss values. This indicates that the spectrograms contain additional information. Additionally, the difference between the training and validation loss for the Spectro-Temporal model is much smaller that the difference between the training and validation loss for the Temporal model.

The discrepancy between the training and validation loss is a key indicator of overfitting. The results show that the

Spectro-Temporal model avoids overfitting better than the Temporal model.

Regarding the evaluation of the overall method, the aggregated results (Table 6) showed that the method achieved an overall accuracy of 92.9% and a score of 89.3, which represents an improvement compared to the results achieved by the *baseline RF* (accuracy of 86.7% and score of 80.2) and *baseline LogReg* (accuracy of 83.1% and score of 70.2). Additionally, our method's accuracy of 92.9% (or the method's error of 7.1%) is quite close to the percentage of recordings labeled as "unknown" by experts (see Table 2), with the rest of the recordings labeled as healthy or unhealthy. This is another indicator that our method performs well for heart-sound classification. Most importantly, our method achieved a 10% higher accuracy compared to the *baseline RF* for the CHF domain (see the results in Table 3 for the dataset U – UKC-JSI).

The windowing analysis in the *UKC-JSI experiments* section showed that for the UKC-JSI dataset, the method achieves an accuracy higher than 81% for all of the tested window sizes. The accuracy achieved by the *baseline RF* on the same dataset was 74.2%, and the accuracy achieved by the *baseline LogReg* was 64.1% (see Table 3). Thus, our method outperforms the baseline methods for any window size, indicating that the method is quite robust with respect to the window size, which was also supported by the LOSO evaluation (see Table 8). It seems that better accuracy is achieved for smaller rather than larger window sizes (see Table 7).

Regarding the related work, one very recent study for measuring the timing of heart sound components through PCG data was presented by Giordano and Knaflitz [50]. Their study focuses on the detection of the typical S1, S2 and S3 heart sounds. A very similar approach was presented by Tseng *et al.* [51]. These approaches depend heavily on the detection of the typical S3 sound, which is mostly conditioned by the quality of the PCG signal. A direct comparison with the related work is presented in Table 11. The baseline methods in this study (*baseline RF* and *baseline LogReg*) use features that are based exactly on the detection of the typical heart sounds [16]. The evaluation results showed that these methods perform poorly compared to our method. The main reason for the poor performance is the noise in the data. More general ML approaches for healthy vs. unhealthy heart sound classification are the methods from the PhysioNet challenge. The winning algorithm of the challenge achieved a score of 85 in their internal evaluation [20]. For the evaluation, they used 20% of the PhysioNet data, which was randomly selected. Our method achieves a score of 89 using a 10-fold cross-validation process, which is basically 10 evaluations over 10% of the data, which was randomly selected. Although the results are not directly comparable, the scores indicate that our method performs quite well compared to the state-of-the-art methods.

Additionally, our main goal was not to create a general method for healthy vs. unhealthy heart-sound detection, but rather, our focus is on CHF detection. In addition to the studies that use the detection of the typical heart sounds (S1, S2, S3), we were not able to find any studies on CHF using machine learning except for our previous work in this field [1], [53]. One very recent study on CHF detection was presented by Porumb et al. [52], where they used CNNs on data collected with ECG devices. Our work differs from theirs, as we used PCG.

TABLE 11. Comparison with the related work.

	Method	Score
Clifford et al. [3]	Logistic Regression (*baseline LogReg*)	72
Clifford et al. [3]	Random Forest (*baseline RF*)	80
Potes et al. [20]	AdaBoost (features) + CNN (end-to-end)	85
Our method	Random Forest (features) + Spectro-temporal ResNet (end-to-end)	89

The feature selection step of the proposed method can significantly influence the classification performance of the feature-based ML methods. In the future, we will conduct an extensive feature selection study to find the optimal feature subset. Additionally, in this work, we tested dataset-specific models. In the future, one could employ transfer-learning techniques [54] to utilize data from other similar studies/datasets. We envision a system capable of detecting new CHF patients by using the proposed method in this study. Once detected, our initial analysis showed that it is relatively simple to build personalized models (e.g., by using some of the identified features in this study) for monitoring the different CHF phases. In the long term, our approach would allow patients to actively cooperate in the process and the treating physicians would be able to adjust patients' medical management in a timely manner and thus avoid hospital admissions.

VII. CONCLUSION

In this paper, we presented a novel method for CHF detection from PCG audio recordings. The method combines classic ML and end-to-end DL. The classical ML learns from a large body of expert-defined features and the DL learns both from the time-domain (i.e., the raw PCG signal) representation of the signal and the spectral representation of the signal. We evaluated the method on our own dataset for CHF detection and additionally on six publicly available PhysioNet datasets used for the recent PhysioNet Cardiology Challenge. The challenge datasets allowed us to extensively evaluate the performance of the method on similar domains. The evaluation results on all the datasets showed that, compared to the challenge baseline methods, our method achieves the best performance (see the *PhysioNet experiments* section). The facts that most of these datasets are labeled for different types of heart-related conditions and that the PCG audio is recorded from a different body position in most of the datasets (e.g., aortic area, pulmonic area, tricuspid area, and mitral area) strongly indicate that the proposed method is quite robust and that it is useful for detecting different

types of heart-sound classification problems and not just for CHF detection, as long as domain-specific labeled data are provided.

Finally, we extended the study beyond the typical healthy vs. patient classification and explored personalized models for detecting different CHF phases, i.e., the recompensated phase (i.e., when the patient feels well) and the decompensated phase (i.e., when the patient needs medical attention). We identified 15 features that have different distributions depending on the phase. By using just two of these features, we were able to build a simple and transparent decision tree classifier (see Fig. 3) that is capable of distinguishing between the recompensated and the decompensated phases with an accuracy of 93.2%, calculated using a LOSO evaluation. While we are aware that there is a risk of overfitting in these final experiments, especially since the dataset contains only 44 samples, we believe that these results are very encouraging and represent a solid base for further development of personalized models. To the best of our knowledge, this is the first study to address such a problem.

ACKNOWLEDGMENT

The authors would like to thank A. Gradišek, MD, who first brought to their attention the idea of detecting CHF based on heart sounds.

REFERENCES

[1] M. Gjoreski, M. Simjanoska, A. Gradisek, A. Peterlin, M. Gams, and G. Poglajen, "Chronic heart failure detection from heart sounds using a stack of machine-learning classifiers," in *Proc. Int. Conf. Intell. Environ. (IE)*, Aug. 2017, pp. 14–19.

[2] J. Voigt, M. S. John, A. Taylor, M. Krucoff, M. R. Reynolds, and C. M. Gibson, "A reevaluation of the costs of heart failure and its implications for allocation of health resources in the United States," *Clin. Cardiol.*, vol. 37, no. 5, pp. 312–321, May 2014.

[3] G. D. Clifford, C. Liu, B. Moody, D. Springer, I. Silva, Q. Li, and R. G. Mark, "Classification of normal/abnormal heart sound recordings: The PhysioNet/computing in cardiology challenge 2016," in *Proc. Comput. Cardiol. Conf. (CinC)*, May 2017, pp. 609–612.

[4] X. Jiang, Y. Pang, X. Li, and J. Pan, "Speed up deep neural network based pedestrian detection by sharing features across multi-scale models," *Neurocomputing*, vol. 185, pp. 163–170, Apr. 2016.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[6] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017.

[7] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018.

[8] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Feb. 2003.

[9] S. Amiriparian, A. Baird, S. Julka, A. Alcorn, S. Ottl, S. Petrovic, E. Ainger, N. Cummins, and B. W. Schuller, "Recognition of echolalic autistic child vocalisations utilising convolutional recurrent neural networks," in *Proc. INTERSPEECH*, Aug. 2018, pp. 2334–2338.

[10] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, S. Pugachevskiy, and B. Schuller, "Bag-of-deep-features: Noise-robust deep feature representations for audio analysis," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–7.

[11] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, "Deep affect recognition from R-R intervals," in *Proc. ACM Int. Joint Conf. Pervas. Ubiquitous Comput. ACM Int. Symp. Wearable Comput. (UbiComp)*, 2017, pp. 754–762.

[12] H. P. Martinez, Y. Bengio, and G. N. Yannakakis, "Learning deep physiological models of affect," *IEEE Comput. Intell. Mag.*, vol. 8, no. 2, pp. 20–33, May 2013.

[13] M. Gams, "Weak intelligence: Through the principle and paradox of multiple knowledge," in *Advances in Computation: Theory and Practice*, vol. 6. Hauppauge, NY, USA: Nova Science Publishers, 2001, p. 245.

[14] M. Gjoreski, V. Janko, G. Slapnicar, M. Mlakar, N. Rescic, J. Bizjak, V. Drobnic, M. Marinko, N. Mlakar, M. Lustrek, and M. Gams, "Classical and deep learning methods for recognizing human activities and modes of transportation with smartphone sensors," *Inf. Fusion.*, 2019.

[15] M. Gjoreski, A. Gradišek, B. Budna, M. Gams, and G. Poglajen, "Toward early detection and monitoring of chronic heart failure using heart sounds," in *Proc. 15th Int. Conf. Intell. Environ. Workshop*, vol. 26. IOS Press, Aug. 2019, p. 336.

[16] S. Schmidt, E. Toft, C. Holst-Hansen, C. Graff, and J. Struijk, "Segmentation of heart sound recordings from an electronic stethoscope by a duration dependent Hidden-Markov model," in *Proc. Comput. Cardiol.*, Sep. 2008, pp. 345–348.

[17] D. B. Springer, L. Tarassenko, and G. D. Clifford, "Logistic regression-HSMM-based heart sound segmentation," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 4, pp. 822–832, Apr. 2015.

[18] M. N. Homsi, N. Medina, M. Hernandez, N. Quintero, G. Perpiñan, A. Quintana, and P. Warrick, "Automatic heart sound recording classification using a nested set of ensemble algorithms," in *Proc. Comput. Cardiol. Conf. (CinC)*, May 2017, pp. 817–820.

[19] F. Plesinger, I. Viscor, J. Halamek, J. Jurco, and P. Jurak, "Heart sounds analysis using probability assessment," *Physiol. Meas.*, vol. 38, no. 8, pp. 1685–1700, May 2017.

[20] C. Potes, S. Parvaneh, A. Rahman, and B. Conroy, "Ensemble of feature: Based and deep learning: Based classifiers for detection of abnormal heart sounds," in *Proc. Comput. Cardiol. Conf. (CinC)*, May 2017, pp. 621–624.

[21] M. Zabihi, A. Bahrami Rad, S. Kiranyaz, M. Gabbouj, and A. K. Katsaggelos, "Heart sound anomaly and quality detection using ensemble of neural networks without segmentation," in *Proc. Comput. Cardiol. Conf. (CinC)*, May 2017, pp. 613–616.

[22] E. Kay and A. Agarwal, "DropConnected neural networks trained on time-frequency and inter-beat features for classifying heart sounds," *Physiol. Meas.*, vol. 38, no. 8, pp. 1645–1657, Jul. 2017.

[23] I. D. Bobillo, "A tensor approach to heart sound classification," in *Proc. Comput. Cardiol. Conf. (CinC)*, May 2017, pp. 629–632.

[24] J. Rubin, R. Abreu, A. Ganguli, S. Nelaturi, I. Matei, and K. Sricharan, "Classifying heart sound recordings using deep convolutional neural networks and mel-frequency cepstral coefficients," in *Proc. Comput. Cardiol. Conf. (CinC)*, May 2017, pp. 813–816.

[25] S. Choi and Z. Jiang, "Comparison of envelope extraction algorithms for cardiac sound signal segmentation," *Expert Syst. Appl.*, vol. 34, no. 2, pp. 1056–1069, Feb. 2008.

[26] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proc. ACM Multimedia (MM)*, vol. 5. Barcelona, Spain: ACM, Oct. 2013, pp. 835–838.

[27] A. Gradišek, G. Slapnicar, J. Šorn, M. Luštrek, M. Gams, and J. Grad, "Predicting species identity of bumblebees through analysis of flight buzzing sounds," *Bioacoustics*, vol. 26, no. 1, pp. 63–76, Jan. 2017.

[28] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011—The first international audio/visual emotion challenge," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.* Berlin, Germany: Springer, Oct. 2011, pp. 415–424.

[29] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.

[30] R. Wang and K. Tang, "Feature selection for maximizing the area under the ROC curve," in *Proc. IEEE Int. Conf. Data Mining Workshops*, Dec. 2009, pp. 400–405.

[31] G. Xuan, X. Zhu, P. Chai, Z. Zhang, Y. Q. Shi, and D. Fu, "Feature selection based on the bhattacharyya distance," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 4, 2006, p. 957.

[32] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information.," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 6, 2004, Art. no. 066138.

[33] C. Arndt, *Information Measures: Information and Its Description in Science and Engineering*. Springer, 2001.

[34] Q. Gu, Z. Li, and J. Han, "Generalized Fisher score for feature selection," 2012, *arXiv:1202.3725*. [Online]. Available: https://arxiv.org/abs/1202.3725

[35] R. F. Woolson, *Wilcoxon Signed-Rank Test*. Hoboken, NJ, USA: Wiley, 2007, pp. 1–3.

[36] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, 1987.

[37] L. Deng and D. Yu, "Deep learning: Methods and applications," *Found. Trends Signal Process.*, vol. 7, nos. 3–4, pp. 197–387, 2014.
[38] Y. Bengio, "Learning deep architectures for AI," *FNT Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.
[39] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.
[40] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural Netw.*, vol. 12, no. 1, pp. 145–151, Jan. 1999.
[41] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proc. 26th Annu. Int. Conf. Mach. Learn. (ICML)*, 2009, pp. 609–616.
[42] K. Jarrett, K. Kavukcuoglu, M. A. Ranzato, and Y. Lecun, "What is the best multi-stage architecture for object recognition?" in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2146–2153.
[43] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[44] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: The difficulty of learning long-term dependencies," Tech. Rep., 2001.
[45] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: https://arxiv.org/abs/1502.03167
[46] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.
[47] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Jan. 2006.
[48] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998.
[49] F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in Statistics*. Springer, 1992, pp. 196–202.
[50] N. Giordano and M. Knaflitz, "A novel method for measuring the timing of heart sound components through digital phonocardiography," *Sensors*, vol. 19, no. 8, p. 1868, Apr. 2019.
[51] Y.-L. Tseng, P.-Y. Ko, and F.-S. Jaw, "Detection of the third and fourth heart sounds using Hilbert-Huang transform," *BioMed. Eng. OnLine*, vol. 11, no. 1, p. 8, 2012.
[52] M. Porumb, E. Iadanza, S. Massaro, and L. Pecchia, "A convolutional neural network approach to detect congestive heart failure," *Biomed. Signal Process. Control*, vol. 55, Jan. 2020, Art. no. 101597.
[53] M. Gjoreski, A. Gradišek, B. Budna, M. Gams, and G. Poglajen, "Toward early detection and monitoring of chronic heart failure using heart sounds," in *Proc. Intell. Environ. Workshop 15th Int. Conf. Intell. Environ.*, 2019, p. 336.
[54] M. Gjoreski, S. Kalabakov, M. Luštrek, and H. Gjoreski, "Cross-dataset deep transfer learning for activity recognition," in *Proc. ACM Int. Joint Conf. Pervas. Ubiquitous Comput. 2019 ACM Int. Symp. Wearable Comput. (UbiComp/ISWC)*, 2019, pp. 714–718.

MARTIN GJORESKI received the B.S. degree in computer science from the Faculty of Computer Science and Engineering, Skopje, Macedonia, in 2014, and the M.S. degree in computer science from the Jozef Stefan Postgraduate School, Ljubljana, Slovenia, in 2016, where he is currently pursuing the Ph.D. degree in computer science.

He has been a Research Assistant with the Department of Intelligent Systems, Jozef Stefan Institute, Ljubljana, since 2014. His research focuses on the development of machine-learning methods for monitoring human physiological and psychological states using wearable sensors. He developed parts of the machine-learning algorithm that received the Sussex-Huawei Locomotion Challenge, in 2018 and 2019.

ANTON GRADI EK received the Ph.D. degree in solid-state physics from the Faculty of Mathematics and Physics, University of Ljubljana, in 2012. He worked as a postdoc at the Korea Basic Science Institute, Daejeon, South Korea, and was a Fulbright Scholar at Washington University, Saint Louis, Missouri. He is currently working at the Jozef Stefan Institute, Ljubljana, Slovenia, at the Department of Intelligent Systems and Department of Solid State Physics. His research interests include hydrogen storage materials, liquid crystals, and applications of artificial intelligence on various domains, especially medicine and bioacoustics. He was a member of a team that reached the finals of the XPrize Tricorder competition.

BORUT BUDNA received the B.Sc. degree in computer and information science from the Faculty of Computer and Information Science, University of Ljubljana, in 2017, where he is currently pursuing the M.Sc. degree. He works as a Student Researcher at the Department of Intelligent Systems, Jožef Stefan Institute. His main research interests include applications of machine learning in medicine and bioacoustics.

MATJA GAMS received the Ph.D. degree. He is currently the Head of the Department of Intelligent Systems, Jožef Stefan Institute, Ljubljana, and is a Professor of computer science with the Jozef Stefan Postgraduate School and University of Ljubljana. His professional interests include intelligent systems, artificial intelligence, cognitive science, intelligent agents, electronic and mobile health, business intelligence, and information society. He is a member of several international program committees of scientific meetings, National and European strategic boards and institutions, and editorial boards of 11 journals, and he is the Managing Director of the journal *Informatica*. He is a member of the National Council of Slovenia, representing the field of science for the term of 2017–2022. His team received two activity recognition competitions and placed in the finals of the XPrize Tricorder competition.

GREGOR POGLAJEN received the M.D. and Ph.D. degrees. He is a Cardiologist staff at the Advanced Heart Failure and Transplantation Center, UMC Ljubljana, Slovenia. He is a member of a focused team, led by prof. Bojan Vrtovec, M.D., Ph.D., whose main clinical interests are advanced heart failure, heart transplantation, and mechanical circulatory support, and his main research focus is in regenerative medicine and remote patient management.