Robust Multimodal Heartbeat Detection Using Hybrid Neural Networks

MICHAEL R. SCHWOB [1], **AEREN DEMPSEY**[1], **FELIX ZHAN**[1], **JUSTIN ZHAN**[2], (Member, IEEE), **AND ASIF MEHMOOD**[3]

[1]University of Nevada, Las Vegas, Las Vegas, NV 89154, USA
[2]University of Arkansas, Fayetteville, AR 72701, USA
[3]Air Force Research Laboratory, Wright-Patterson Air Force Base, OH 45433, USA

Corresponding author: Justin Zhan (jzhan@uark.edu)

ABSTRACT Many arrhythmia datasets are multimodal due to the simultaneous collection of physiological signals of a subject. These datasets frequently have missing modalities or missing block-wise data, a characteristic that various recent applications of neural networks fail to consider. Most arrhythmic detection models only use electrocardiogram and blood pressure recordings. Unconsidered physiological signals may be strongly correlated with other modalities despite having missing data. To improve robustness and accuracy of heartbeat detection, all available modalities should be considered in multimodal arrhythmia datasets. Several hybrid neural networks are proposed to robustly analyze heartbeats by considering every available physiological signal. These networks combine elements from convolutional neural networks, recurrent neural networks, and a deep learning architecture. This enables researchers to analyze every signal of subjects while the set of signals collected among subjects may differ. The proposed hybrid neural networks provide more robust results in heartbeat detection when utilizing missing data modalities.

INDEX TERMS Multimodal, heartbeat detection, deep learning, neural networks.

I. INTRODUCTION

Biological events can be documented by multiple signals, or modalities. If multiple modalities are recorded for an event, then the existing multimodal data may reveal characteristics that each modality might not independently uncover. In recent years, much attention has been placed on analyzing multimodal data through machine learning [1].

Many multiparameter datasets have incomplete information in the form of missing modalities and missing block-wise data. This can occur when modalities are recording data at different times or when sensors fail to record data under certain conditions. When some modalities have more complete data than others, an asymmetric multimodal dataset is obtained. Many recent methods fail to obtain accurate detection results from these asymmetric datasets, since they do not account for the missing data.

Multimodal medical imaging datasets frequently have missing data. Physicians may not record every physiological signal for every patient. Some subjects may have respiratory

rates recorded while others will have electrocardiogram (ECG) recordings. After appending a group of patients' data together, some modalities may have missing data. For example, if three patients have ECG recordings, while four other patients do not, an appended matrix will reveal missing data for the ECG modality. If neural networks train on large datasets where observations may have unique combinations of various biometric recordings, then these neural networks must be able to analyze modalities that have missing data. Various mechanisms and data processing techniques have been utilized to robustly analyze multimodal datasets with missing data: late fusion, co-learning, orthogonal regularization, probabilistic graphic models, and deep Boltzmann machines.

In recent years, various deep learning methods have been proposed. More frequently, neural networks, including convolutional neural networks (CNNs) [2], recurrent neural networks (RNNs) [3], and modular neural networks [4], have been used to analyze multimodal data. These artificial neural networks, inspired by biological neural networks, learn to perform tasks through a training dataset. Once trained, the neural networks should accurately classify inputs

into various output categories and can be used to analyze biometrics.

The entirety of a multimodal arrhythmia dataset should be considered to robustly detect heartbeats. Although some modalities may contain missing data, or some modalities may be missing in many observations, these physiological signals may have a strong correlation with other signals. Alternatively, some signals may be too noisy and decrease heartbeat detection accuracy. The intrinsic structure of an arrhythmia dataset would only be revealed through an analysis of every modality.

Three contributions are provided in this paper. First, several hybrid neural networks (HNNs) are proposed to analyze multiple modalities while obtaining robust results for heartbeat detection. Second, an evaluation of the proposed HNNs on three real-world multimodal arrhythmia datasets is provided. Third, a comparison of CNN, CNN-Dropout, long short-term memory (LSTM), gated recurrent unit (GRU), EmbraceNet (EN), and the proposed HNNs is provided, along with a discussion regarding which common physiological signals should and should not be included in future analyses.

In Section II, related research to biometric anomaly analysis, heartbeat detection, and deep learning with neural networks is explored. The discussed research will be related to neural network theory and robust multimodal heartbeat detection. In Section III, multiple HNNs are proposed to improve the robustness of heartbeat detection with missing multimodal data. In Section IV, the experimental set-up, datasets, and experimental results are reported. In Section V, a conclusion discusses the findings, and possible future work is presented.

II. RELATED RESEARCH

Anomaly detection has been applied to many biological domains [5]–[10]. Several papers propose the implementation of anomaly detection within existing medical wireless sensor networks [8], [9], while others propose to detect outliers with medical devices [11], [12]. Since anomalies within biological systems may be fatal, much attention has been focused on the detection of system abnormalities.

Deep learning has been widely applied to many applications, from agriculture [13] to medical image analysis [14], [15]. In recent years, various deep learning methods have been proposed to analyze large datasets [16]–[18]. With more advanced computing devices and larger datasets available, it became possible to utilize deep learning architectures with a large number of layers [19], [20].

A. DEEP LEARNING FOR MULTIMODAL DATASETS

Often times, several biometric parameters are recorded simultaneously for an individual subject, generating a multimodal dataset. Many methods have been proposed to apply neural networks to multimodal biometric datasets [4], [21]–[28], several of which attempted to increase robustness for multimodal datasets with missing data [29]–[31].

A multimodal ensemble-based system for emotional detection has revealed robust qualities when exploring fusion methods for emotional recognition [29]. This technique solves the missing data problem at the multimodal fusion stage.

Deep Boltzmann machines are stochastic RNNs that have been proposed for learning a generative model for multimodal datasets [30]. Since most of these datasets are images, CNNs are favored over RNNs for detecting biometric anomalies, a noteworthy example being arrhythmia [32]. LSTMs have produced significant results in time-series anomaly detection [33], as well as multimodal biometric anomaly detection [34]. Although GRUs have not seen many multimodal applications, it may prove to be advantageous for smaller datasets [35].

EN, a deep learning architecture, was designed for multimodal tasks that consider cross-modal correlations while avoiding over-fitting. Missing information in one modality can be covered by other signals, revealing some robustness when encountering missing modalities or block-wise missing data [31]. Although EN reveals robust qualities when analyzing multimodal datasets with missing data, the conventional CNN demonstrates superior feature extraction for images, while taking advantage of local spatial coherence [36].

B. DEEP LEARNING FOR HEARTBEAT DETECTION

Various neural network types have been utilized in several disciplines. Recent progress has frequently been made in classification and detection problems [37], [38], most prominently in heartbeat detection. Many methods are used with artificial neural networks, including multi-domain feature extraction [36], backpropagation [39], [40], and linear and nonlinear feature combination [41].

Substantial work has also been done in heartbeat detection using CNNs. One significant example is found in CNN-based generalized information fusion, which reveals the implications of using multi-channel data to increase accuracy [42]. Directly reading the multimodal data, similar to what would be found in hospital readings, mitigates the need for intermediate estimating methods like signal differencing, filter-banks, wavelet transform, and Hilbert transform [43]–[46]. Examples of other solutions include the use of co-occurrence matrices [47] and 3-dimensional data structures [48].

Some efforts have been made to apply a hybrid CNN-LSTM neural network to detect biometric anomalies, such as arrhythmia [32], [49], [50], or to detect diseases, such as diabetes [51]. Neither of these efforts attempt to apply a CNN-LSTM architecture to a multimodal dataset. However in the field of gesture recognition [52], CNN-LSTM architectures were applied to multimodal datasets.

C. DATA PRE-PROCESSING

A need for data pre-processing arises due to the existence of noisy readings, which limits the efficiency of deep learning architectures [53]. A multitude of solutions have been proposed, including combining finite impulse response (FIR)

filters and principal component analysis (PCA) [54], wavelet transform approaches [55]–[57], modified empirical mode decomposition [58], and iterative denoising [59].

The main objective of combining FIR filters and PCA is to minimize base line wandering, which significantly improves accuracy. Alternatively, wavelet transform approaches aim to change the wavelet coefficients rather than the actual base line. However, in these approaches, the scalability is limited. In an attempt to mitigate this, the multiple wavelets transform theorem, based on the ideas of the single wavelet transform and multi-resolution analysis, was proposed, showing improvement upon the traditional single-wavelet transform method.

Similarly, the modified empirical mode decomposition method combines the empirical mode decomposition method and the nonlocal means method to conserve the morphological characteristics of the wavelets, while maintaining high accuracy.

In an effort to maximize overall accuracy, an iterative denoising method was implemented, applying every wavelet function and possible decomposition to denoise the wavelets. Consequently, accuracy increased at the cost of computation speed.

III. APPROACH

To propose novel HNNs, a mathematical framework must be provided to support the integration of the different architectures. These percentages are calculated following a comparison of existing neural networks and the proposed HNNs, one can determine the deep learning architectures that best handle arrhythmia datasets with missing modalities or missing block-wise data.

A. DATA PROCESSING

Although the proposed HNNs can be implemented in most biometric applications, it is critical to correctly pre-process the data. A generalized pre-processing technique is provided for heartbeat analysis.

Since the physiological signals are obtained as time-series data, non-overlapping windows of the samples are scaled by the maximum magnitude. Additionally, normalizing physiological data increases learning speed and decreases convergence time. Missing values were filled with the mean of their neighboring values. All signals were resampled to 250 Hz and normalized to a range of [0,1].

Each HNN is optimized with its weights throughout the training and validation phases using cross-validation for each dataset. In an effort to provide the most rigorous analysis of the proposed HNNs, cross validation is implemented on the training set.

Snippets are generated from the biometric samples by moving a window of l length one sample at a time, where l can be selected based on the physiological signals analyzed. The value of l can be generalized to a length of 251 for most arrhythmia detection applications. The generated snippets are used to train and optimize the presented neural architectures.

B. ACTIVATION FUNCTIONS

Most recent attempts to determine heartbeat location use the logistic sigmoid activation function $\sigma(x) = \frac{1}{1+e^{-x}}$, which outputs data between [0,1] [42], [60]. This activation function's derivative, $\frac{d\sigma}{dx} = \frac{e^{-x}}{(1+e^{-x})^2}$, centers the data at 0. However, the *tanh* activation function, $tanh = 2\sigma(2x) - 1$, avoids bias in convolutional gradients. Additionally, *tanh* accomplishes the same goal as the logistic sigmoid, since $\frac{d(tanh)}{dx} = \frac{4e^{-2x}}{(e^{-2x}+1)^2}$. The rectified linear units (ReLU) activation function, $f(x) = \max(0, x)$, and leaky ReLU activation function, $g(x) = \max(\alpha x, x)$, have gained attention in recent research, where α is a small constant (i.e. 0.1). These four activation functions are used throughout the experiment.

Although previous attempts at detecting heartbeats utilize a logistic sigmoid nonlinearity in the fully connected network, the hyperbolic tangent nonlinearity, ReLU, or leaky ReLU are tested with the proposed HNNs. The sigmoid nonlinearity is not sensitive with values close to 0 or 1. Additionally, it is hard to initialize weights of sigmoid neurons to prevent saturation. If weights are slightly large, the neural network may have difficulties learning. Finally, the sigmoid activation is not zero-centered, which is critical if a neural network will operate on more than one convolution layer.

C. CONVOLUTIONAL NEURAL NETWORKS

Simultaneously, K physiological signals are recorded for a subject from time t_1 to time t_2, denoted $x^{(i)}_{t_1:t_2}$ where $1 \leq i \leq K$. Since multimodal datasets are analyzed, $K \geq 2$.

Like many neural networks, CNNs can systematically fuse information from multiple signals by adding convolution layers between the input layer and the output layer. With K signals, each filter contains K one-dimensional component filters with a common length L. The desired feature set, or overall filter output, is generated by summing these component filter outputs. By assuming p filters, the coefficients of the k-th channel of the j-th filter is denoted $\{h_l^{(k,j)}\}_{l=0}^{L-1}$, where $1 \leq k \leq K$ and $1 \leq j \leq p$. The corresponding filter output is denoted $y_n^{(k,j)} = \Sigma_{l=1}^{L-1} x_{n-l}^{(k)} h_l^{(k,j)}$. By summing all channels, the overall output of the j-th filter is obtained: $y_n^{(j)} = \Sigma_{k=1}^{K} y_n^{(k,j)}$.

At a given start time t, a window of length M of the signal vector $x^{(k)}_{t:t+M-1}$ for the k-th channel is considered. Without zero-padding, a relatively short filter can produce a larger number of outputs than a long filter, which would improve accuracy for smaller datasets. To avoid zero-padding, the j-th filter output $y_n^{(j)}$ is obtained for $n = t+L-1, \cdots, t+M-1$. This results in $M-L+1$ samples. The length of the produced feature vector for all filters is $p(M-L+1)$.

The filter coefficients and network weights were optimized using back propagation under the cross-entropy loss function

$$\mathscr{L}(X, Y) = -\frac{1}{n}\Sigma_{i=1}^{n} y^{(i)} \ln a(x^{(i)}) + (1 - y^{(i)}) \ln(1 - a(x^{(i)}))$$

(1)

where $X = \{x^{(1)}, \cdots, x^{(n)}\}$ is the set of input examples in the training dataset, $Y = \{y^{(1)}, \cdots, y^{(n)}\}$ is the corresponding label set for those input examples, and $a(x)$ is the output of a given input x. Throughout the experiments, label 1 will be used for inputs that are determined a heartbeat and the label 0 will denote the absence of a heartbeat. The values of $a(x)$ are computed through one of the following functions:

$$a_1(x) = \frac{1}{1 + e^{-(Wx+b)}} \tag{2}$$

$$a_2(x) = \frac{2}{1 + e^{-2(Wx+b)}} - 1, \tag{3}$$

$$a_3(x) = \max(0, Wx + b), \tag{4}$$

$$a_4(x) = \max(\alpha x, Wx + b), \tag{5}$$

where W is a weight matrix, b is a bias vector, and α is a small, positive constant. The functions $a_1(x)$, $a_2(x)$, $a_3(x)$, and $a_4(x)$ are the logistic sigmoid, hyperbolic tangent, ReLU, and leaky ReLU activation functions, respectively. Since many arrhythmia datasets contain more than two physiological signals, the proposed HNNs may use multiple layers. If two layers are used, the activation functions can be expanded to

$$a_1(x) = \frac{1}{1 + e^{-(Wz_1(x)+b)}}, \tag{6}$$

$$a_2(x) = \frac{2}{1 + e^{-2(Wz_2(x)+b)}} - 1, \tag{7}$$

$$a_3(x) = \max(0, Wz_3(x) + b), \tag{8}$$

$$a_4(x) = \max(\alpha z_4(x), Wz_4(x) + b), \tag{9}$$

where $z_1(x) = \frac{1}{1+e^{-(Vx+c)}}$, $z_2(x) = \frac{2}{1+e^{-2(Vx+c)}} - 1$, $z_3(x) = \max(0, Vx+c)$, $z_4 = \max(\alpha x, Vx+c)$, V is the weight matrix for the first layer, c is the bias vector for the first layer, and $z_i(x)$ for $i = 1, 2, 3, 4$ is the activation layer of the hidden layer. If multiple hidden layers exist, this process is repeated.

D. LONG SHORT-TERM MEMORY
One of the more popular RNNs for biometric analysis is the LSTM. This model is trained to reconstruct the normal time-series data inherent in the physiological recordings. Any reconstruction errors are used to obtain the likelihood of a point denoting a heartbeat. Since LSTM units do not perform activation, the same physiological signals can flow through the neural network for an arbitrarily long period of time. This characteristic enables LSTMs to represent time series, where the interaction between past and present data is sensitive to distant and recent events [61].

The jth unit of a LSTM maintains a memory c_t^j at time t. The output of the jth unit is $h_t^j = o_t^j \sigma(c_t^j)$, where o_t^j is an output gate modulating the amount of content exposure for the memory and $\sigma(\cdot)$ is an activation function. Given a sequence $x = (x_1, x_2, \cdots, x_r)$, this gate is computed as

$$o_t^j = \sigma(W_o x_t + U_o h_{t-1} + V_o c_t)^j, \tag{10}$$

where W and U are weight matrices, h_t is the model's recurrent hidden state at time t, and V_o is a diagonal matrix.

The memory c_t^j is updated by adding a temporary memory $c_t^{j\prime}$ after partially forgetting the existing memory, where $c_t^{j\prime} = \sigma(W_c x_t + U_c h_{t-1})^j$. The extent of the temporary loss of the existing memory is controlled by a forget gate $f_t^j = \sigma(W_f x_t + U_f h_{t-1} + V_f c_{t-1})^j$. The extent to which the new memory content is added to c_t^j is modulated by an input gate $i_t^j = \sigma(W_i x_t + U_i h_{t-1} + V_i c_{t-1})^j$. Similar to the output gate, V_f and V_i are diagonal matrices. An LSTM unit uses these gates to determine whether the existing or new memory content carries more information in feature extraction. If an LSTM model detects an important feature early in the training stage, its unit carries this information throughout the training, thereby capturing potential long-distance dependencies.

E. GATED RECURRENT UNIT
In recent years, GRUs, a modified LSTM, have been proposed, which allows each unit to adaptively capture dependencies on different time scales [62]. GRUs do not separate memory cells when modulating the flow of information inside each recurrent unit.

When working on less training data, GRUs train faster and perform better on testing data than LSTMs. Due to the simplicity of GRUs, they compute much more efficiently [63]. However, LSTMs can remember longer sequences than GRUs and more accurately model long-distance relationships [61], [64].

For a GRU, the activation $h_t^j = (1 - z_t^j)h_{t-1}^j + z_t^j h_t^{j\prime}$ acts as a linear interpolation between the previous activation h_{t-1}^j and the new activation $h_t^{j\prime} = \sigma(Wx_t + U(r_t \odot h_{t-1}))^j$, where r_t is a set of reset gates. The update gate $z_t^j = \sigma(W_z x_t + U_z h_{t-1})^j$ determines how much of the unit is updated.

When $r_t^j \approx 0$, the reset gate forces the recurrent unit to forget the previously computed state. Similar to the update gate, $r_t^j = \sigma(W_r x_t + U_r h_{t-1})^j$.

One key difference between the LSTM and GRU unit is that the GRU unit does not have a mechanism to control the degree of exposure for the unit's state; instead, the GRU unit exposes the state entirely at any given time.

F. EmbraceNet
The deep learning architecture EN has the robust quality of dealing with missing data [31]. Assuming there exists K physiological signals with corresponding neural networks, let $x^{(k)}$ be the output vector from the k-th neural network, where $k \in \{1, 2, \cdots, K\}$. Through the use of *docking layers*, EN converts each input vector into vectors of the same size. The i-th component of the k-th docking layer is denoted $z_i^{(k)} = w_i^{(k)} \cdot x^{(k)} + b_i^{(k)}$, where $w_i^{(k)}$ is a weight vector and $b_i^{(k)}$ is a bias vector. An activation function is applied to $z_i^{(k)}$ to obtain the output of the k-th docking layer, $d_i^{(k)} = f_a(z_i^{(k)})$, where $d^{(k)} = [d_1^{(1)}, d_2^{(1)}, \cdots, d_c^{(k)}]$ and c is the dimension of all docked vectors. With the K vectors of dimension c obtained from the docking layers, EN combines these vectors into an *embraced vector*.

The fusion technique of the K docking vectors uses a multinomial sampling, where $r_i = [r_i^{(1)}, r_i^{(2)}, \cdots, r_i^{(K)}]^T$ is a vector drawn from a multinomial distribution (i.e. $r_i \sim$ Multinomial$(1, p)$), where $p = [p_1, p_2, \cdots, p_m]^T$ and $\Sigma_k p_k = 1$. From these constraints, one value of r_i will equal 1 with the remaining values equal to 0. Thereafter, the vector $d'^{(k)} = [d_1'^{(k)}, d_2'^{(k)}, \cdots, d_c'^{(k)}]^T = r^{(k)} \circ d^{(k)}$ is obtained from a Hadamard product, where $r^{(k)} = [r_1^{(k)}, r_2^{(k)}, \cdots, r_c^{(k)}]^T$. The embracement layer $e = [e_1, e_2, \cdots, e_c]^T$ is calculated such that $e_i = \Sigma_k d_i'^{(k)}$. This procedure is outlined in Algorithm 1.


Algorithm 1 EmbraceNet

Require: m neural network output vectors, f_a
1: $x^{(k)} := k$th model output vector, where $1 \leq k \leq m$
2: $w_i^{(k)} :=$ weight vector for the ith component of the kth model
3: $b_i^{(k)} :=$ bias vector for the ith component of the kth model
4: $f_a :=$ non-linearity
5: **for** $1 \leq i \leq c$ **do**
6: $\quad z_i^{(k)} = w_i^{(k)} \cdot x^{(k)} + b_i^{(k)}$
7: $\quad d_i^{(k)} = f_a(z_i^{(k)})$
8: **end for**
9: $d^{(k)} = [d_1^{(k)}, d_2^{(k)}, \cdots, d_c^{(k)}]^T$
10: **for** $1 \leq i \leq c$ **do**
11: $\quad r_i \sim$ Multinomial$(1, p)$
12: **end for**
13: $d'^{(k)} = r^{(k)} \circ d^{(k)}$
14: **for** $1 \leq i \leq c$ **do**
15: $\quad e_i = \Sigma_k d_i'^{(k)}$
16: **end for**
17: **return** $e = [e_1, e_2, \cdots, e_c]^T$

This process ensures that one physiological signal contributes to each component of the vector e. Through the modality selection process, the output of EN is generated from data of all signals. The selection process depends on the values of p. In cases where little to no data is missing, $p = [1/K, 1/K, \cdots, 1/K]^T$ gives an equal chance for all physiological signals to be selected.

However, arrhythmia datasets often have missing modalities or missing block-wise data. In this case, the probabilities p are adjusted. Let $u = [u_1, u_2, \cdots, u_K]^T$ indicate the presence of each signal, where $u_k = 1$ if $x^{(k)}$ exists and 0 otherwise. The multinomial distribution is adjusted by changing p to $\hat{p} = [\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_K]^T$, where $\hat{p}_k = \frac{u_k p_k}{\Sigma_j u_j p_j}$. If the k-th signal is not available, $u_k = 0$ and $\hat{p}_k = 0$, eliminating the chance for that value of $r_i^{(k)}$ to become 1. This prevents invalid data coming from the k-th physiological signal to propagate to the EN output.

G. HYBRID NEURAL NETWORKS

Recent attempts at detecting heartbeats make use of two physiological signals, or channels: ECGs and blood pressure (BP). An analysis of these two modalities is a bimodal approach. The proposed HNNs can accurately analyze many channels.
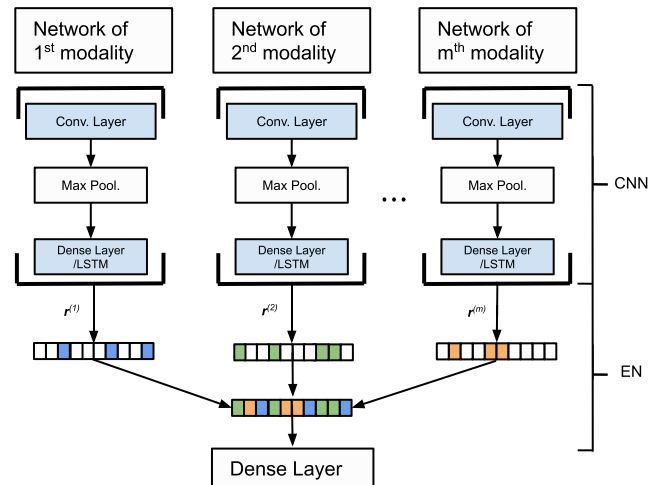
FIGURE 1. Diagram of the proposed CNN-LSTM-EN neural architecture.

Since the proposed architectures are not restricted to unimodal or bimodal datasets, they may contain more than one convolution layer. Additionally, many arrhythmia datasets contain missing data, which deters a multimodal ($K > 2$) approach to analyze the data.

A combination of neural network architectures may reveal a more robust approach to analyzing datasets that have missing data for heartbeat detection; due to the rhythmic nature of this detection and the necessity for quality feature extraction with missing data, a single neural network may not be advantageous. It is for this reason that many researchers are implementing HNNs like CNN-LSTM and CNN-GRU in multimodal datasets.

CNNs are ideal when data is periodically sampled in one or more dimensions, which is the most common occurrence for arrhythmia data. LSTMs are useful when the data display a periodic rhythm and can return accurate results when detecting heartbeats. GRUs provide the benefits of LSTMs on smaller datasets. The benefit of EN arises when datasets have missing modalities or missing block-wise data, which is a common occurrence for the volatile nature of physiological recordings. Therefore, combinations of these four neural architectures may provide accurate, robust results. HNNs can incorporate many channel signals by CNN, LSTM, or GRU and feed these channels into the EN for final prediction. A diagram of the proposed CNN-LSTM-EN neural network is provided in Figure 1.

An explanation of the CNN-LSTM-EN is provided first, since it is the most involved neural network alongside the CNN-GRU-EN architecture. Note that since a GRU is a modified LSTM, the CNN-GRU-EN is similarly constructed to the CNN-LSTM-EN.

Since CNNs read periodically sampled data well, this neural network will be the first architecture in the proposed CNN-LSTM-EN model, leveraging its superior feature extraction over the other neural architectures. The benefit of including an LSTM in an HNN is that LSTMs can best analyze periodic data. EN is the last architecture since it

requires periodically analyzed data to adjust the probabilities $\hat{p}$ before attempting to detect heartbeats in the arrhythmia data. This same reasoning justifies the structure of the existing hybrid network CNN-LSTM, where the LSTM is the architecture that classifies heartbeat location rather than an EN.

Typically, a CNN's probability vector p_{cnn} is generated prior to data classification. In the CNN-LSTM-EN neural architecture, the dense layer of the ith modality is an LSTM, where $i \in \{1, 2, \cdots, m\}$. Following the feature extraction of the convolution layers in the m CNN architectures, the LSTM will produce a probability vector p_l that determines the probabilities of points denoting heartbeats, where $l \in \{1, 2, \cdots, m\}$. Rather than executing the classification process after the dense layer, p_l will become the input to the EN. An EN will take the m probability vectors as the input vectors, where the docking layers are computed as

$$z_l^{(k)} = w_l^{(k)} \cdot p^{(k)} + b_i^{(k)}. \tag{11}$$

Thereafter, the EN architecture will compute more robust results by adapting the probabilistic distributions of the available data and then classifying heartbeat locations.

For the CNN-LSTM and CNN-LSTM-EN neural networks, the CNN architecture conducts the feature extraction. The extracted features become the input for the LSTM model. This is advantageous since the CNN extracts features from the arrhythmia data with accuracy while the LSTM detects arrhythmic anomalies following the CNN's feature extraction. The CNN-EN architecture operates similarly to the CNN-LSTM-EN, where the m probability vectors of the CNN models become the input vectors for the EN architecture.

As a modified LSTM, GRUs can also detect arrhythmic conditions following the implementation of a CNN. Therefore, CNN-GRU and CNN-GRU-EN models are examined throughout the experiment. Although LSTM-EN and GRU-EN models cannot leverage the periodic sampling technique and feature extraction of CNNs, their results are included in the experiment as a means of comparison.

CNN, regularized CNN with dropout, GRU, LSTM, and EN are included in the experiment as a baseline for comparing the results between single neural architectures and their hybrid counterparts. In total, twelve neural architectures are analyzed.

H. FEATURE EXTRACTION
Compared to scattering networks, RNNs, and EN, CNNs boast several additional strengths: (1) there are a wide variety of filters (random, supervised, and unsupervised filters) that can be employed (2) there exists a variety of non-linearities (rectified linear units, hyperbolic tangent, and logistic sigmoid) (3) pooling separators (sub-sampling, average-pooling, and max-pooling) can be applied and (4) these filters, non-linearities, and pooling separators can be in different network layers. With these additional features, CNNs have more customization and flexibility than scattering networks, RNNs, and EN.

Let $L^p(\mathbb{R}^d)$ be the space of Lebesgue-measurable functions $f : \mathbb{R}^d \to \mathbb{C}$ that satisfies $||f||_p := (\int_{\mathbb{R}^d} |f(x)|^p dx)^{1/p} < \infty$, where $p \in [1, \infty)$. Since semi-discrete frames can be interpreted as shift-invariant frames of a countable index Λ_n with a continuous translation parameter, the following definitions can be provided.

In the nth network layer for a CNN, a convolution with atoms $g_{\lambda_n} \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ of a semi-discrete Parseval frame $\Psi_n := \{T_b I_{g_{\lambda_n}}\}_{b \in \mathbb{R}^d, \lambda_n \in \Lambda_n}$ for $L^2(\mathbb{R}^d)$ is employed on a countable index set Λ_n, where T_b and $I_{g_{\lambda_n}}$ are frame coefficients outlined in [65] and the frame atoms g_{λ_n} are arbitrary. The semi-discrete frame Ψ_n functions as a feature extractor.

For $n \in \mathbb{N}$, let M_n and P_n be Lipschitz-continuous operators where $M_n f = P_n f = 0$ for $f = 0$. Using the previously defined semi-discrete frame Ψ_n, the sequence $\Omega = ((\Psi_n, M_n, P_n))_{n \in \mathbb{N}}$ is called a module-sequence.

Given the module-sequence $\Omega = ((\Psi_n, M_n, P_n))_{n \in \mathbb{N}}$, the feature extractor Ψ_n, with Ω, maps $L^2(\mathbb{R}^d)$ to the feature vector

$$\Phi_\Omega(f) := \bigcup_{n=0}^{\infty} \Phi_\Omega^n(f), \tag{12}$$

where $\Phi_\Omega^n(f) := \{(U[q]f) * \chi_n\}_{q \in \Lambda_1^n}, \forall n \in \mathbb{N}, U[\cdot]$ is the operator associated with the nth network layer, and the function χ_n is the output-generating atom of the nth layer. The set $\Phi_\Omega^n(f)$ corresponds to the features in the nth network layer generated by function f.

The feature extractors obtained through the implemented neural architectures use this feature-extracting mechanism to store all features in the feature vectors $\Phi_{\Omega_i}(f)$ before classifying the biometric data analyzed.

IV. EXPERIMENT
The experiments were run with a NVIDIA Tesla P100, a 16-GB GPU computing processor. The NVIDIA Tesla P100 contains 3584 CUDA cores with a bandwidth of 720 GBps. The experimental results follow a description of the used datasets.

A. DATASETS
To ensure robustness of the proposed HNNs, three multimodal datasets were analyzed: the MIT-BIH polysomnographic dataset, the MIT CC original dataset, and the MIT CC augmented dataset. These datasets include a variety of sample lengths. In each dataset, the patient record contains notes from cardiologists on the heartbeat locations. More specifically, the weighted mean of the R-peak and S-peak of an ECG is described as the heartbeat location. The individual databases are expanded upon below.

  1) **MIT-BIH Polysomnographic Database**
     The MIT-BIH dataset contains four-, six-, and seven-channel polysomnographic recordings that sum to over 80 hours worth of data [66]. Additionally, heartbeat annotations, ECGs, EEGs, and respiration signals
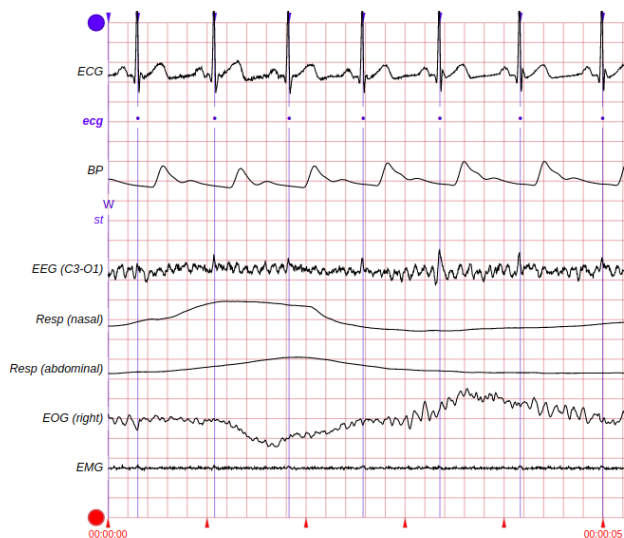
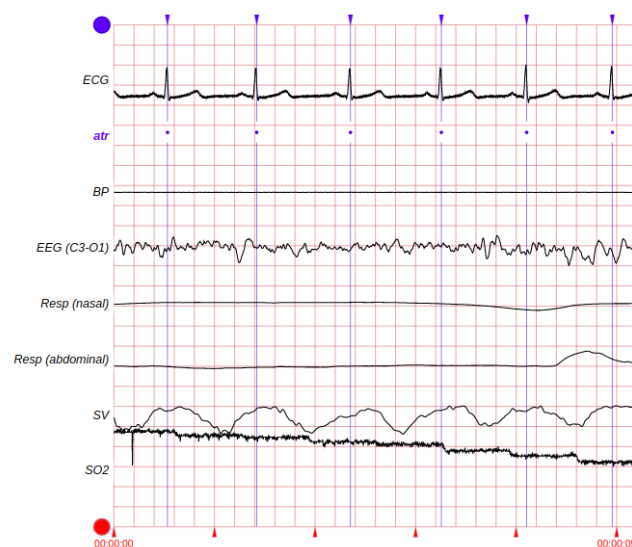**FIGURE 2.** Sample 45 in the MIT-BIH polysomnographic database.

FIGURE 3. Sample 100 in the MIT CC database.

annotated with respect to sleep conditions are provided. These multimodal recordings were obtained during sleep for eighteen subjects. A LightWAVE visualization of subject 45 from the MIT-BIH polysomnographic dataset is provided [Fig. 2]. This sample was selected to display the variety in available modalities for each sample.

2) **MIT Computing in Cardiology (CC) Database**

The publicly available MIT CC database contains 200 patient records and a hidden test set with 210 records [67]. This database contains two datasets: an original dataset and an augmented dataset. Both datasets contain 100 training observations. Since the augmented dataset does not have a fixed fiducial point and the annotations were not generated from a specific physiological channel, this dataset provides a greater challenge for the neural networks. Signals' sampling frequencies vary between 250 and 360 Hz.

A LightWAVE visualization of sample 100 from the MIT CC database is provided [Fig. 3]. This sample was selected to display the variety in available modalities for each sample within the multimodal database. Physician annotations that are recorded in blue indicate a heartbeat. *ECG, BP, EEG, SV,* and *SO2* represent the physiological signals for ECGs, BP, electroencephalography, stroke volume and sulfur dioxide, respectively. *Resp (nasal)* and *Resp (abdominal)* indicate signals for nasal respiration and abdominal respiration, respectively.

B. EXPERIMENTAL RESULTS

The testing accuracies for the CNN, CNN-Dropout, CNN-LSTM, and the proposed HNNs across fifty epochs for all three datasets are displayed in Figures 4, 5, and 6. These figures are included to reveal the effect that modality count
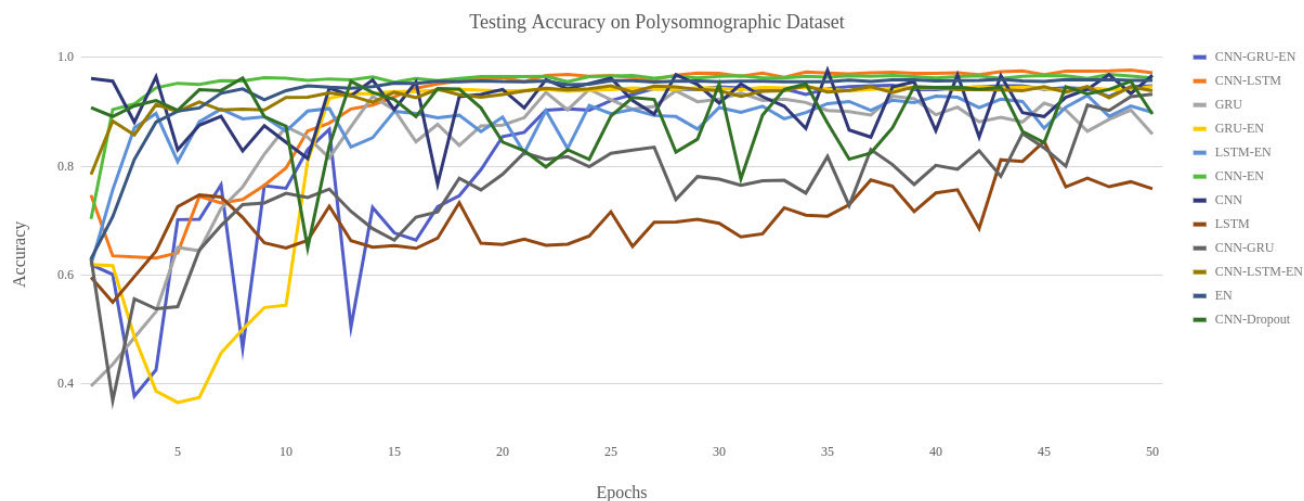

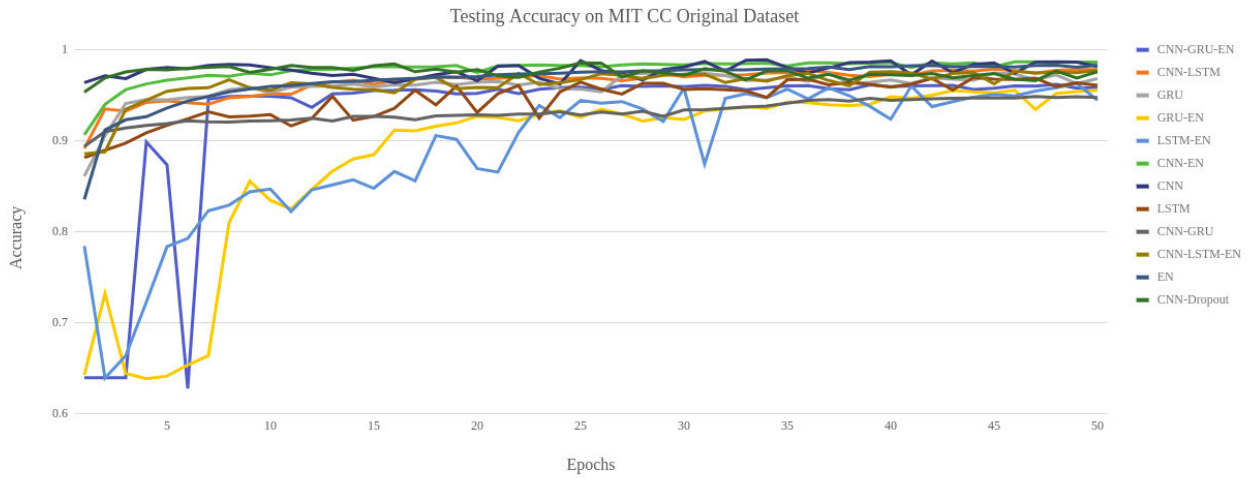**FIGURE 4.** Testing accuracy on the MIT polysomnographic dataset.

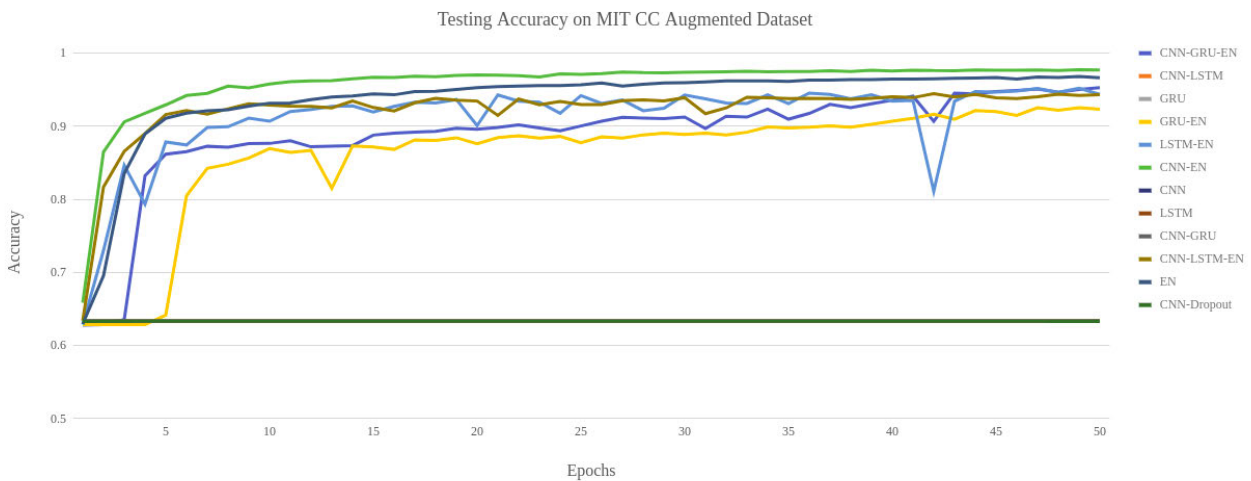**FIGURE 5.** Testing accuracy on the MIT CC original dataset.

FIGURE 6. Testing accuracy on the MIT CC augmented dataset.

has on the learning of each implemented neural network. When analyzing the testing accuracies for these neural networks, we compare only their best performance. The testing accuracies are reported as percentages following each model for the different datasets, where testing accuracy is defined as the percentage of correctly detected heartbeats. A comparison is made between each neural network's detected heartbeats and the physician-annotated heartbeats in each dataset.

Tables 1, 2, and 3 report the change in percentage points for each model after dropping each modality in the three datasets. In some cases, the performance of a model improves when particular modalities are removed. This suggests that several modalities contribute noisy and uncorrelated data, since removing their impact on the network increases that network's predictability. However, several models reveal that common physiological signals can improve the accuracy of locating heartbeats. In Table 3, neural architectures that did not experience a change in percentage points for dropped modalities were removed.

The neural networks that reported the best accuracy on the MIT-BIH Polysomnographic database contained convolution layers. Namely, the CNN (97.63%), CNN-LSTM (97.62%), CNN-EN (96.86%), and CNN-Dropout (96.19%) models reported the highest accuracies on the polysomnographic testing data. This indicates that the MIT-BIH data requires more advanced feature extraction to obtain accurate results, which is best satisfied with neural architectures that leverage the feature extraction of CNNs. Note that despite CNN-LSTM having a higher accuracy than CNN at most epochs, CNN obtains the highest accuracy at epoch 35. Although the best testing accuracies were reported by the CNN and CNN-LSTM architectures, these models contained high variance in accuracy when modalities were dropped, indicating a relatively heavy reliance on ECG and BP readings. Additionally, the regularized CNN-Dropout model displayed relatively high variance.

The CNN-EN model had comparable results with the CNN and CNN-LSTM in terms of the best reported accuracy while exhibiting less volatility. This is likely due to the

TABLE 1. Change in percentage points for dropped modalities on the MIT polysomnographic dataset.

Dropped Modality	CNN	Dropout[1]	GRU	LSTM	CNN-GRU	CNN-LSTM	EN	CNN-EN	GRU-EN	LSTM-EN	C-G-E[2]	C-L-E[3]
(Original Accuracy)	97.63	96.19	85.90	75.86	93.20	97.62	95.90	96.86	94.59	89.95	94.87	94.76
ECG	-50.9	-36.3	-49.0	-38.9	0.8	-34.1	-1.4	-1.9	-0.5	0.9	0.0	0.1
BP	-35.3	-25.5	-22.8	-12.8	-30.1	-34.1	-6.8	-0.3	-22.5	-32.0	-42.2	-35.4
Resp (nasal)	-0.7	-1.4	1.2	-0.4	0.0	0.0	-0.1	0.0	-0.2	0.0	0.0	0.1
EEG (C3-O1)	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	-0.1	-0.1	-0.2	-0.1	-0.1
Resp (abdominal)	0.0	0.0	0.0	0.0	0.0	0.0	-0.2	0.0	-0.1	-0.1	0.0	0.0
EEG (C4-A1)	-0.6	5.9	-10.6	1.0	1.6	0.0	0.0	0.0	0.0	0.3	0.0	0.0
EMG	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	0.2	0.0	-0.2
SV	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	-0.1	0.0	0.0	-0.1
SO2	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	-0.1	-0.1	-0.2	0.0	0.0
EOG (right)	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	-0.1	0.0	-0.1	0.0	0.0
Resp (chest)	0.0	0.0	0.0	0.0	0.0	0.0	-0.2	-0.1	-0.1	-0.1	0.0	-0.1
EEG (O2-A1)	-0.3	0.3	2.0	11.4	0.0	0.0	-0.2	-0.2	-0.1	0.0	0.0	0.0
Resp (sum)	0.0	5.2	5.9	-2.8	3.0	-1.5	-0.1	-0.1	-0.1	0.4	0.0	-0.2
EOG	0.0	0.0	0.0	0.0	0.0	0.0	-0.2	-0.1	-0.1	-0.2	0.0	-0.1
Resp (abdomen)	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	-0.1	-0.1	-0.1	0.0	0.0

TABLE 2. Change in percentage points for dropped modalities on the MIT CC original dataset.

Dropped Modality	CNN	Dropout	GRU	LSTM	CNN-GRU	CNN-LSTM	EN	CNN-EN	GRU-EN	LSTM-EN	C-G-E	C-L-E
(Original Accuracy)	98.91	98.56	96.82	96.09	94.86	97.82	98.31	98.68	95.56	94.47	95.83	97.72
ECG	-7.9	-9.9	-3.1	-21.6	-1.0	-7.1	-3.8	-4.0	-0.7	-3.3	-37.4	-3.6
BP	-12.0	-18.7	-29.2	-31.8	-22.5	-14.0	-2.0	0.0	-17.7	-5.5	0.2	-0.8
Resp (nasal)	-0.3	0.0	-0.2	-1.4	-0.1	0.0	0.1	0.0	0.1	-0.1	0.1	0.1
EEG (C3-O1)	-0.1	-0.1	-0.3	-0.3	-0.1	0.0	0.0	0.0	0.2	0.0	0.1	0.0
Resp (abdominal)	0.1	0.0	0.0	-0.1	0.1	-0.1	0.1	0.0	0.0	0.0	0.2	0.0
EEG (C4-A1)	-0.2	0.0	-0.5	-0.6	0.0	0.0	0.0	0.1	0.1	0.2	0.2	0.0
EMG	-0.2	0.0	-0.1	-0.2	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0
SV	0.4	-0.2	0.6	0.8	0.0	0.1	0.1	0.1	0.0	-0.2	0.1	0.1
SO2	-0.1	-0.2	-0.1	-0.7	-5.0	-0.5	0.0	0.0	0.0	0.2	0.1	0.1
EOG (right)	0.0	0.0	0.1	0.1	0.0	-0.1	0.0	0.0	0.0	0.1	0.0	-0.1
Resp (chest)	0.0	-0.1	-0.4	-0.2	-0.3	0.0	0.0	-0.1	0.2	0.2	0.0	0.0
EEG (O2-A1)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Resp (sum)	0.5	1.2	-0.2	0.1	0.1	0.2	0.0	0.0	0.1	0.1	0.2	0.1
EOG	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	0.0	0.1	0.0	0.0
Resp (abdomen)	0.0	-0.1	-0.2	-0.3	-0.2	0.0	0.1	-0.1	0.1	0.1	0.0	0.0

EN's ability to adjust probability assessments while working with the CNN, which would prove advantageous since the MIT-BIH data requires a more rigorous treatment towards feature extraction than analyzing a periodic rhythm. When analyzing arrhythmia datasets, a high testing accuracy with low variance is ideal; therefore, the results of CNN-EN are comparable to the CNN and CNN-LSTM models, while CNN-EN retains higher testing accuracy after modalities are removed. The GRU-EN (94.59%), CNN-GRU-EN (94.87%), and CNN-LSTM-EN (94.76%) models also reported stable testing accuracies following the removal of modalities.

The MIT CC original dataset reported similar results to the polysomnographic dataset. The CNN (98.91%), CNN-EN (98.68%), CNN-Dropout (98.56%), and EN (98.31%) models recorded the highest testing accuracy. Similar to the MIT-BIH data, this dataset requires advanced feature extraction. CNN-EN demonstrates less variance in testing accuracy than a regularized CNN with comparable results, likely due to the ability to adjust the probability vector outputted from the CNN architecture. The models containing RNNs such as the GRU-EN (95.56%) and CNN-GRU (94.86%) models performed relatively worse than the other HNNs, indicating that the MIT CC original dataset does not require a rigorous periodic analysis of the physiological signals.

The most accurate models for the MIT CC augmented dataset utilized an embracement layer for late fusion. The CNN-EN (97.74%), EN (96.83%), CNN-GRU-EN (95.28%), and LSTM-EN (95.23%) reported the best results, since the EN architecture works well with a large number of modalities (46 physiological signals). The neural architectures without an EN layer failed to learn after the first epoch, since the relatively large amount of modalities inhibited learning; each of these neural architectures reported a 63.34% testing accuracy throughout the experiment, only utilizing ECG and BP readings as inputs.

For the three datasets tested, ECG and BP recordings were the most common physiological signals and have the largest impact on the outcomes of the models. Tables 1, 2, and 3 report the change in percentage points of accuracy as certain modalities are dropped from the input to the neural networks.

For the MIT polysomnographic dataset, the HNNs that do not utilize EN were mostly impacted by the removal of the ECG modality. After removing ECG signals from the observations, the CNN, GRU, LSTM, CNN-Dropout, and CNN-LSTM lost 51, 49, 39, 36, and 34 percentage points of accuracy, respectively. The CNN-EN, GRU-EN, and EN only lost 2, 1, and 1 percentage points, respectively, following the removal of ECG signals. The neural architectures

TABLE 3. Change in percentage points for dropped modalities on the MIT CC augmented dataset.

Dropped Modality	EN	CNN-EN	GRU-EN	LSTM-EN	C-G-E	C-L-E
(Original Accuracy)	96.83	97.74	92.51	95.23	95.28	94.26
BP	-5.3	-0.2	-28.6	-15.2	-2.2	-31.4
ECG	-2.3	-4.3	0.7	-0.5	-2.2	0.2
Resp (nasal)	0.1	0.1	0.6	0.0	0.1	0.2
No signal	0.1	0.0	-0.1	0.0	0.0	0.0
EEG (C3-O1)	0.2	0.1	0.4	0.1	0.0	0.2
Pressure	0.1	0.0	0.1	-0.1	0.1	0.0
ART	0.1	0.0	0.0	0.0	0.0	0.0
CVP	0.0	0.0	0.0	0.0	0.2	0.0
PAP	0.0	0.0	0.0	-0.2	0.1	0.0
ECG lead II	0.0	0.0	0.1	0.0	0.1	0.0
EMG	0.2	0.0	0.5	0.0	0.0	0.1
Resp (abdominal)	0.1	0.2	0.3	0.0	0.1	0.1
Resp. Imp.	0.1	0.1	0.0	0.0	0.1	0.0
CO2	0.0	0.1	0.0	0.0	0.1	-0.1
EOG (right)	0.1	0.1	0.4	0.1	0.2	0.1
EEG (C4-A1)	0.2	0.1	0.2	0.0	0.1	0.2
SV	0.2	0.1	0.1	0.0	0.1	0.1
SO2	0.0	0.0	0.2	0.1	0.1	0.1
Resp (chest)	0.1	0.0	0.0	-0.1	0.0	0.0
ECG Lead II	0.0	0.0	0.0	0.0	-0.1	0.0
Nothing	0.1	0.0	0.0	0.0	0.0	0.0
ECG II	0.1	0.1	0.2	0.1	0.1	0.1
Pressure 2	0.0	0.1	0.0	-0.1	0.1	0.0
EEG (O2-A1)	0.1	0.0	0.0	0.1	0.0	-0.1
EOG	0.1	-0.1	0.0	0.0	0.1	0.1
Resp (abdomen)	0.1	0.1	0.1	0.1	0.1	0.0
Resp (sum)	0.2	0.0	0.0	-0.3	0.2	0.0
Pressure 1	0.1	0.1	-0.1	-0.2	0.0	0.1
ECG III	0.2	0.0	0.0	-0.1	0.0	-0.1
Pressure 3	0.1	0.0	0.1	-0.1	0.0	0.0
II	0.0	0.1	0.0	0.0	0.2	0.1
Pressure1	0.1	0.0	-0.1	0.0	0.1	0.0
ECG Lead AVF	0.1	0.0	0.1	0.1	0.1	0.0
CO2 off	0.1	0.0	-0.1	-0.1	0.2	-0.1
Pressure 4	0.1	0.0	-0.1	0.0	0.1	0.1
Resp.Imp.	0.1	0.1	-0.1	-0.1	0.0	0.1
PAP off	0.0	0.0	-0.1	0.0	0.0	0.0
CVP off	0.0	0.0	0.0	-0.1	0.0	0.0
ECG lead 2	0.0	0.0	-0.1	0.1	0.0	0.0
ART 1	0.0	0.0	0.0	-0.1	0.0	0.0
PAP 2	0.0	-0.1	0.0	-0.1	0.1	0.0
CVP 3	0.1	0.1	-0.1	-0.1	0.1	0.1
ECG Lead III	0.1	0.0	0.1	0.0	0.2	0.0
No PAP	0.1	0.0	0.1	-0.2	0.0	0.0
ECG lead AVL	-0.1	0.0	0.1	0.0	0.1	0.0
ECG 2	0.0	0.0	0.0	-0.1	0.1	0.1

without an embracement layer lost the most percentage points in accuracy because they failed to make use of the other physiological signals. With EN, the proposed HNNs were able to remain relatively accurate following the removal of either the ECG or BP modality. The CNN-LSTM-EN and CNN-GRU-EN architectures report little to no loss in percentage points for accuracy. The CNN-GRU and LSTM-EN models report a 1 percentage point increase in accuracy after removing ECG, indicating that these models better predict heartbeat locations from BP than ECG recordings. These findings indicate that the proposed HNNs and EN outperform the state-of-the-art neural networks when common modalities are not present, displaying the robustness of HNNs.

When BP signals are removed, CNN-EN barely loses accuracy, whereas CNN, CNN-LSTM, CNN-GRU,

CNN-Dropout, GRU, and LSTM lose 35, 34, 30, 26, 23, and 13 percentage points, respectively. Additionally, CNN-GRU-EN, CNN-LSTM-EN, LSTM-EN, GRU-EN, and EN experience a loss of 42, 35, 32, 22, and 7 percentage points, respectively. The CNN-EN architecture remains relatively accurate compared to the CNN-LSTM-EN and CNN-GRU-EN architectures following the removal of the BP modality because this dataset does not require a rigorous periodic analysis to be completed by the RNNs. None of the HNNs that incorporate an embracement layer lose any accuracy when dropping physiological signals other than BP or ECG recordings, since they are able to make use of each modality in detecting heartbeats. For the polysomnographic dataset, CNN-EN provides comparable testing accuracy to CNN, regularized CNN and CNN-LSTM while providing much more robust accuracy following the

removal of modalities. Therefore, in a messy arrhythmia dataset, CNN-EN provides robust results while offering testing accuracy that competes well with state-of-the-art neural networks.

Within the MIT polysomnographic dataset, ECG, BP, and nasal respiration appear to positively affect the networks' testing accuracies, whereas EEG (O2-A1) and the summation of respiration appear to negatively affect testing accuracies. While most of the remaining physiological signals do not greatly impact neural network performance, EEG (C4-A1) provides either a positive or negative impact on testing accuracies depending on the tested neural architecture.

The MIT CC original dataset reports similar results to the polysomnographic dataset. While CNN, CNN-EN, CNN-Dropout, and EN provide the best testing accuracies, EN and CNN-EN provide the most robust results when dropping modalities. The HNNs that implement an embracement layer do not lose accuracy when dropping more of the uncommon physiological signals, as well. However, GRU-EN and CNN-GRU-EN lose a significant amount of accuracy after dropping BP or ECG signals, since the MIT CC original dataset has longer periods of recordings that the GRU architecture struggles to analyze. Thus, EN and CNN-EN report some of the highest testing accuracies while providing modality robustness.

For the MIT CC original dataset, the neural architectures that use an embracement layer tend to perform better after removing modalities, with the exception of BP and ECGs. However, the neural networks without EN reported lower testing accuracies following the removal of most modalities. Thus, the proposed HNNs that utilize an embracement layer report robust results following the removal of both common and uncommon biometric signals. Within the MIT CC original dataset, ECG, BP, nasal respiration, EEG, sulfur dioxide, chest respiration, and abdomen respiration appear to positively affect the networks' testing accuracies. Stroke volume, EOG (right), and the summation of respiration appear to negatively affect testing accuracies.

Since the neural architectures that did not incorporate EN did not learn on the MIT CC augmented dataset, dropping any modality would not affect their testing accuracy. Therefore, their results report a 0% change and are excluded from Table 3. It is likely that these neural architectures did not learn because they only analyzed the ECG and BP modalities, which did not reveal additional information past the first epoch. Relative to the other datasets, the CNN-EN, EN, and CNN-GRU-EN performed well on the MIT CC augmented dataset, only losing 4, 2, and 2 percentage points after dropping ECG signals, respectively. This is likely due to the relatively heavier reliance on the EN architecture over the CNNs and RNNs, since this dataset contains a larger amount of modalities. Since the removal of the ECG modality negatively impacts the accuracy of almost every neural network on each dataset, ECG is a strong indicator for heartbeat location; thus, ECGs should be considered in multimodal heartbeat detection when available.

Both LSTM-EN and CNN-LSTM-EN did not experience a change in accuracy after dropping ECG signals, revealing that they make use of some of the rather uncommon physiological signals. Therefore, if ECG recordings experience block-wise missing data in arrhythmia datasets that contain a large amount of modalities, the proposed LSTM-EN and CNN-LSTM-EN neural networks provide accurate and robust results. The GRU-EN model suffers a loss of 29 percentage points following the removal of BP recordings while improving by 1 percentage point after dropping ECG signals. This finding indicates that the GRU-EN model relied more on BP than ECG recordings to locate heartbeats.

Additionally, GRU-EN was the only model that gained any percentage points in accuracy after dropping a physiological signal that was not BP or ECG, which can be attributed to the inherent noise in some of the more uncommon physiological signals. Based on testing accuracy and modality robustness, CNN-EN, EN, and CNN-GRU-EN appear to best predict heartbeat location on the MIT CC augmented dataset.

On the augmented dataset, BP provides a strong positive impact on the networks' testing accuracies, whereas ECG affects accuracies relatively less. Several physiological signals negatively affect network performance for each neural architecture: nasal respiration, EEG (C3-O1), EMG, abdominal respiration, EOG (right), EEG (C4-A1), stroke volume, sulfur dioxide, ECG II, abdomen respiration, and ECG Lead AVF. The remaining physiological signals have either little or no effect on testing accuracies. In the two other datasets, nasal respiration, EEG (C3-O1), and abdomen respiration positively affected testing accuracies. This discrepancy may be due to the amount of noise inherent in this dataset since there are more biometric signals present.

V. CONCLUSION AND FUTURE WORK

Many biometric datasets contain more than two physiological signals. To make full use of each dataset, all modalities must be considered. As seen in the experimental results, not every modality will positively contribute to predictive accuracy. However, some modalities, which have been neglected up until this point by use of bimodal analysis, positively contribute to the predictive accuracy of many commonly used neural networks.

This finding reveals room for improvement in many state-of-the-art multimodal approaches that only analyze two modalities. Regarding arrhythmia studies, most researchers only analyze BP and ECGs. However, other physiological signals such as nasal respiration, EEGs, and sulfur dioxide can positively contribute to predictive accuracy when implementing several neural architectures.

As indicated throughout Tables 1-3, the proposed HNNs do not necessarily outperform the state-of-the-art neural networks when analyzing complete datasets, although the HNNs do report similar accuracies. However once confronted with missing data or modalities, the commonly used CNN, LSTM, CNN-LSTM, and GRU lose a significant amount of accuracy.

In comparison, the proposed HNNs report robust results, losing a much smaller amount of accuracy with datasets containing missing data.

The CNN-EN provides the most robust testing accuracy of all of the models implemented on the MIT Polysomnographic Database. Since EN is capable of robustly handling missing data and CNNs read periodically sampled data very well, the CNN-EN borrows the strengths from each individual architecture. This model performed particularly well on messier, noisier data compared to the other tested architectures. For cleaner, complete datasets, such as the MIT CC original dataset, CNN with regularization appears to provide sufficient accuracy; however, CNN-EN boasted similar testing accuracy, while losing minimal accuracy following dropped modalities.

For the MIT CC original dataset, EN and CNN-EN provide higher accuracies while mitigating the volatility of their testing accuracies. Additionally, these models experience a minimal drop in accuracy after removing common and uncommon physiological signals. Therefore, EN and CNN-EN are ideal for arrhythmia datasets that are sparse or abundant in modalities.

Since the MIT CC augmented dataset contained up to 46 modalities for each subject, most models failed to learn. However, the HNNs that incorporated EN continued to learn throughout the fifty epochs of the experiment. EN, CNN-EN, and CNN-GRU-EN provided the most robust results, while providing high accuracies for heartbeat locations.

Future research directions may include implementing a multimodal fusion architecture prior to feature extraction for any of the tested neural architectures, through the use of adaptive, sensor, tensor, or memory fusion networks. Additionally, one can further test the efficacy of the proposed HNNs on other biometric datasets with a variety in size and dimensionality.

It is common that a patient has multiple missing modalities. Therefore, an experiment analyzing the robustness of the state-of-the-art neural architectures and the proposed HNNs on biometric datasets when multiple modalities are dropped may reveal more interesting findings.

Since the EN and CNN-EN neural networks can be implemented on messier datasets with missing modalities, previously collected databases that were overlooked due to missing data can be revisited to gain new insights.

If one could identify missing block-wise data and report these locations to the HNNs, testing accuracies may improve. By telling the HNNs that incorporate EN which modalities are missing, testing accuracy may further improve. A more computationally expensive method for improving accuracy could be individually training a neural network on each modality and using their outputs as inputs for an embraced layer.

REFERENCES

[1] T. Baltrusaitis, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[2] Y. Wei, "Cross-modal retrieval with CNN visual features: A new baseline," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 449–460, Feb. 2017.

[3] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," 2014, *arXiv:1412.6632*. [Online]. Available: http://arxiv.org/abs/1412.6632

[4] D. Hidalgo, "Type-1 and type-2 fuzzy inference systems as integration methods in modular neural networks for multimodal biometry and its optimization with genetic algorithms," *Inf. Sci.*, vol. 179, no. 13, pp. 2123–2145, 2009.

[5] S. Chauhan and L. Vig, "Anomaly detection in ECG time signals via deep long short-term memory networks," in *Proc. IEEE Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2015, pp. 1–7.

[6] W. R. Arney and B. J. Bergen, "The anomaly, the chronic patient and the play of medical power," *Sociol. Health Illness*, vol. 5, no. 1, pp. 1–24, 1983.

[7] R. Stafford, "Application of neural networks as an aid in medical diagnosis and general anomaly detection," Tech. Rep., Jul. 1994.

[8] O. Salem, A. Guerassimov, A. Mehaoua, A. Marcus, and B. Furht, "Sensor fault and patient anomaly detection and classification in medical wireless sensor networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2013, pp. 4373–4378.

[9] O. Salem, A. Guerassimov, A. Mehaoua, A. Marcus, and B. Furht, "Anomaly detection in medical wireless sensor networks using SVM and linear regression models," *Int. J. E-Health Med. Commun.*, vol. 5, no. 1, pp. 20–45, Jan. 2014.

[10] A. Burner, "Texture bags: Anomaly retrieval in medical images based on local 3D-texture similarity," in *Proc. MICCAI Int. Workshop Med. Content-Based Retr. Clin. Decis. Support*. Springer, 2011, pp. 116–127.

[11] M. Zhang, A. Raghunathan, and N. K. Jha, "MedMon: Securing medical devices through wireless monitoring and anomaly detection," *IEEE Trans. Biomed. Circuits Syst.*, vol. 7, no. 6, pp. 871–881, Dec. 2013.

[12] N. Jha et al., "Securing medical devices through wireless monitoring and anomaly detection," U.S. Patent 10 135 849 B2, Nov. 2, 2018.

[13] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Comput. Electron. Agricult.*, vol. 147, no. 2, pp. 70–90, Apr. 2018.

[14] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[15] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[16] L. Deng, "A tutorial survey of architectures, algorithms, and applications for deep learning," *APSIPA Trans. Signal Inf. Process.*, vol. 3, no. 1, pp. 1–29, Jan. 2014.

[17] Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "A survey on deep learning for big data," *Inf. Fusion*, vol. 42, pp. 146–157, Jul. 2018.

[18] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," *ACM Comput. Surveys*, vol. 52, no. 1, pp. 1–38, Feb. 2019.

[19] Y. LeCun, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[20] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434

[21] S. Das and M. K. Kundu, "NSCT-based multimodal medical image fusion using pulse-coupled neural network and modified spatial frequency," *Med. Biol. Eng. Comput.*, vol. 50, no. 10, pp. 1105–1114, Oct. 2012.

[22] C.-H. Chen and C. Chu, "Fusion of face and iris features for multimodal biometrics," in *Proc. Int. Conf. Biometrics*. Springer, 2006, pp. 571–580.

[23] A. Gavrilova and M. Monwar, *Multimodal Biometrics and Intelligent Image Processing for Security Systems*. 2013.

[24] Y.-P. Wang, "Multimodal medical image fusion using fuzzy radial basis function neural networks," in *Proc. Int. Conf. Wavelet Anal. Pattern Recognit.*, vol. 2, 2007, pp. 778–782.

[25] Z. Wang and Y. Ma, "Medical image fusion using m-PCNN," *Inf. Fusion*, vol. 9, no. 2, pp. 176–185, Apr. 2008.

[26] X. Xu, D. Shan, G. Wang, and X. Jiang, "Multimodal medical image fusion using PCNN optimized by the QPSO algorithm," *Appl. Soft Comput.*, vol. 46, pp. 588–595, Sep. 2016.

[27] Z. Liu, H. Yin, Y. Chai, and S. X. Yang, "A novel approach for multimodal medical image fusion," *Expert Syst. Appl.*, vol. 41, no. 16, pp. 7425–7435, Nov. 2014.

[28] S. Das and M. K. Kundu, "Ripplet based multimodality medical image fusion using pulse-coupled neural network and modified spatial frequency," in *Proc. Int. Conf. Recent Trends Inf. Syst.*, Dec. 2011, pp. 229–234.

[29] J. Wagner, E. Andre, F. Lingenfelser, and J. Kim, "Exploring fusion methods for multimodal emotion recognition with missing data," *IEEE Trans. Affect. Comput.*, vol. 2, no. 4, pp. 206–218, Oct. 2011.

[30] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2222–2230.

[31] J.-H. Choi and J.-S. Lee, "EmbraceNet: A robust deep learning architecture for multimodal classification," *Inf. Fusion*, vol. 51, pp. 259–270, Nov. 2019.

[32] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ECG classification by 1-D convolutional neural networks," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 664–675, Mar. 2016.

[33] P. Malhotra, "Long short term memory networks for anomaly detection in time series," Presses universitaires de Louvain, Louvain, Belgium, Tech. Rep., 2015, p. 89.

[34] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G. Shroff, "LSTM-based encoder-decoder for multi-sensor anomaly detection," 2016, *arXiv:1607.00148*. [Online]. Available: http://arxiv.org/abs/1607.00148

[35] Y. Yuan, C. Tian, and X. Lu, "Auxiliary loss multimodal GRU model in audio-visual speech recognition," *IEEE Access*, vol. 6, pp. 5573–5583, 2018.

[36] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep CNN features," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5012–5024, Nov. 2016.

[37] T. W. Cenggoro, A. H. Kridalaksana, E. Arriyanti, and M. I. Ukkas, "Recognition of a human behavior pattern in paper rock scissor game using backpropagation artificial neural network method," in *Proc. 2nd Int. Conf. Inf. Commun. Technol. (ICoICT)*, May 2014, pp. 238–243.

[38] M. Soranamageswari and C. Meena, "An efficient feature extraction method for classification of image spam using artificial neural networks," in *Proc. Int. Conf. Data Storage Data Eng.*, Feb. 2010, pp. 169–172.

[39] E. Al-Masri, "Detecting ECG heartbeat abnormalities using artificial neural networks," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 5279–5281.

[40] R. Ghongade and A. A. Ghatol, "A brief performance evaluation of ECG feature extraction techniques for artificial neural network based classification," in *Proc. IEEE Region 10 Conf. (TENCON)*, Oct. 2007, pp. 1–4.

[41] F. A. Elhaj, N. Salim, A. R. Harris, T. T. Swee, and T. Ahmed, "Arrhythmia recognition and classification using combined linear and nonlinear features of ECG signals," *Comput. Methods Programs Biomed.*, vol. 127, pp. 52–63, Apr. 2016, doi: 10.1016/j.cmpb.2015.12.024.

[42] B. S. Chandra, C. S. Sastry, and S. Jana, "Robust heartbeat detection from multimodal data via CNN-based generalizable information fusion," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 3, pp. 710–717, Mar. 2019.

[43] P. Phukpattaranont, "QRS detection algorithm based on the quadratic filter," *Expert Syst. Appl.*, vol. 42, no. 11, pp. 4867–4877, Jul. 2015, doi: 10.1016/j.eswa.2015.02.012.

[44] F. Zhang and Y. Lian, "QRS detection based on multiscale mathematical morphology for wearable ECG devices in body area networks," *IEEE Trans. Biomed. Circuits Syst.*, vol. 3, no. 4, pp. 220–228, Aug. 2009.

[45] I. I. Christov, "Real time electrocardiogram qrs detection using combined adaptive threshold," *Biomed. Eng. OnLine*, vol. 3, no. 1, p. 28, 2004, doi: 10.1186/1475-925X-3-28.

[46] J. Rubin, R. Abreu, A. Ganguli, S. Nelaturi, I. Matei, and K. Sricharan, "Recognizing abnormal heart sounds using deep learning," Tech. Rep., 2017.

[47] W. Sun, N. Zeng, and Y. He, "Morphological arrhythmia automated diagnosis method using gray-level co-occurrence matrix enhanced convolutional neural network," *IEEE Access*, vol. 7, pp. 67123–67129, 2019.

[48] F. Li, Y. Xu, Z. Chen, and Z. Liu, "Automated heartbeat classification using 3-D inputs based on convolutional neural network with multi-fields of view," *IEEE Access*, vol. 7, pp. 76295–76304, 2019.

[49] S. L. Oh, E. Y. K. Ng, R. S. Tan, and U. R. Acharya, "Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats," *Comput. Biol. Med.*, vol. 102, pp. 278–287, Nov. 2018.

[50] Ö. Yildirim, "A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification," *Comput. Biol. Med.*, vol. 96, pp. 189–202, May 2018.

[51] G. Swapna, S. Kp, and R. Vinayakumar, "Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals," *Procedia Comput. Sci.*, vol. 132, pp. 1253–1262, 2018.

[52] G. Zhu, L. Zhang, P. Shen, and J. Song, "Multimodal gesture recognition using 3-D convolution and convolutional LSTM," *IEEE Access*, vol. 5, pp. 4517–4524, 2017.

[53] S. M. Khorasani, G. A. Hodtani, and M. M. Kakhki, "Effects of pre-processing on the ECG signal sparsity and compression quality," in *Proc. 8th Int. Conf. Comput. Knowl. Eng. (ICCKE)*, Oct. 2018, pp. 268–273.

[54] R. V. Savitha, S. R. Breesha, and X. F. Joseph, "Pre processing the abdominal ECG signal using combination of FIR filter and principal component analysis," in *Proc. Int. Conf. Circuits, Power Comput. Technol. (ICCPCT)*, Mar. 2015, pp. 1–4.

[55] E. M. El Hassan and M. Karim, "An FPGA-based implementation of a pre-processing stage for ECG signal analysis using DWT," in *Proc. 2nd World Conf. Complex Syst. (WCCS)*, Nov. 2014, pp. 649–654.

[56] W. Yan, C. Xu, J. Li, and J. Huang, "The use of discrete wavelets for signal denoising in microchip capillary electrophoresis," in *Proc. 6th World Congr. Intell. Control Autom.*, 2006, pp. 5254–5258.

[57] S. Gang, R. Liang, and G. Yuhan, "Research on the method of ECG denoising based on improved wavelet packet," in *Proc. IEEE Int. Conf. Cyber Technol. Autom., Control, Intell. Syst. (CYBER)*, Jun. 2015, pp. 1425–1428.

[58] P. Singh, S. Shahnawazuddin, and G. Pradhan, "Significance of modified empirical mode decomposition for ECG denoising," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2017, pp. 2956–2959.

[59] H. G. Rodney Tan, A. C. Tan, P. Y. Khong, and V. H. Mok, "Best wavelet function identification system for ECG signal denoise applications," in *Proc. Int. Conf. Intell. Adv. Syst.*, Nov. 2007, pp. 631–634.

[60] Y. Özbay and G. Tezel, "A new method for classification of ECG arrhythmias using neural network with adaptive activation function," *Digit. Signal Process.*, vol. 20, no. 4, pp. 1040–1049, Jul. 2010.

[61] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*. [Online]. Available: http://arxiv.org/abs/1412.3555

[62] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," 2014, *arXiv:1409.1259*. [Online]. Available: http://arxiv.org/abs/1409.1259

[63] Ł. Kaiser and I. Sutskever, "Neural GPUs learn algorithms," 2015, *arXiv:1511.08228*. [Online]. Available: http://arxiv.org/abs/1511.08228

[64] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2342–2350.

[65] T. Wiatowski and H. Bolcskei, "A mathematical theory of deep convolutional neural networks for feature extraction," *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1845–1866, Mar. 2018.

[66] Y. Ichimaru and G. B. Moody, "Development of the polysomnographic database on CD-ROM," *Psychiatry Clin. Neurosciences*, vol. 53, no. 2, pp. 175–177, Apr. 1999.

[67] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000.

MICHAEL R. SCHWOB is currently pursuing the bachelor's degree in mathematics with the University of Nevada, Las Vegas. Due to his diverse research experience in astrophysics, bioinformatics, big data, and economics, he was named as 2019 Goldwater Scholar. As a Research Assistant with the Big Data Hub, University of Nevada, Las Vegas, he explores his research interests in spatio-temporal statistics, Bayesian analysis, game theory, biostatistics, hypergraph theory, machine learning, and ecological modeling.

AEREN DEMPSEY is currently pursuing the bachelor's degree in computer science with the University of Nevada, Las Vegas. He is also a Research Assistant with the Big Data Hub, University of Nevada, Las Vegas. His main research focuses on machine learning, data science, and deep learning.

FELIX ZHAN is currently pursuing the degree with the Ed W. Clark High School, Las Vegas, NV, USA. He is also a non-degree seeking student at the University of Nevada, Las Vegas. He has various work and volunteer experiences, including being the President and a Student Tutor of the Computer Science Club at his high school, a Volunteer and a Competitor for Science Olympiad, a Summer Intern at the U.S. Air Force Research Laboratory and Wright State University, a Program Manager, Programming Instructor, Research Mentor, and Website Developer for the UNITE-Army Educational Outreach Program at the University of Nevada, Las Vegas, and a Volunteer of Las Vegas Music Teacher Association. His research interests include artificial intelligence, big data analytics, deep learning, cyber security, and social computing. He has published 15 articles in international journals and conferences, including IEEE ACCESS and *Journal of Big Data* and the IEEE INTERNATIONAL CONFERENCE ON INFORMATION REUSE AND INTEGRATION FOR DATA SCIENCE, the IEEE INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND KNOWLEDGE ENGINEERING, the IEEE INTERNATIONAL CONFERENCE ON BIG KNOWLEDGE, and the IEEE ANNUAL CONSUMER COMMUNICATIONS AND NETWORKING CONFERENCE.

JUSTIN ZHAN (Member, IEEE) is an ARA Scholar and a Professor of data science at the Department of Computer Science and Computer Engineering, College of Engineering, University of Arkansas. He is also a Professor at the Department of Biomedical Informatics, College of Medicine, University of Arkansas for Medical Sciences. His research interests include data science, biomedical informatics, artificial intelligence, information assurance, and social computing. He has published 240 articles in peer-reviewed journals and conferences and delivered more than 30 keynote speeches and invited talks. He has been involved in more than 50 projects, as a principal investigator (PI) or a Co-PI, which were funded by the National Science Foundation, Department of Defense, and National Institute of Health. He was the Steering Chair of the IEEE International Conference on Social Computing (SocialCom), the IEEE International Conference on Privacy, Security, Risk and Trust (PASSAT), and the IEEE International Conference on BioMedical Computing (BioMedCom). He has been the Editor-in-Chief of the *International Journal of Information Privacy, Security and Integrity* and *International Journal of Social Computing and Cyber-Physical Systems*. He has served as the conference general chair, the program chair, the publicity chair, the workshop chair, and a program committee member for 150 international conferences; he has also served as the editor-in-chief, an editor, an associate editor, a guest editor, an editorial advisory board member, and an editorial board member for 30 journals.

ASIF MEHMOOD is a Senior Research Engineer at Air Force Research Laboratory (AFRL). He serves as a senior technical specialist and a researcher with extensive experience in developing and implementing algorithms in signal processing, pattern recognition, image processing, and machine learning. His primary research focuses on target detection, pattern recognition, target tracking, and classification. He is currently working on several deep learning-based projects, such as super resolution style transfer employing generative adversarial networks.

● ● ●