MHDeep: Mental Health Disorder Detection System based on Wearable Sensors and Artificial Neural Networks

SHAYAN HASSANTABAR, JOE ZHANG, HONGXU YIN, AND NIRAJ K. JHA, Princeton University, USA

Mental health problems impact the quality of life of millions of people around the world. However, diagnosis of mental health disorders is a challenging problem that often relies on self-reporting by patients about their behavioral patterns and social interactions. Therefore, there is a need for new strategies for diagnosis and daily monitoring of mental health conditions. The recent introduction of body-area networks consisting of a plethora of accurate sensors embedded in smartwatches and smartphones and edge-compatible deep neural networks (DNNs) points towards a possible solution. Such wearable medical sensors (WMSs) enable continuous monitoring of physiological signals in a passive and non-invasive manner. However, disease diagnosis based on WMSs and DNNs, and their deployment on edge devices, such as smartphones, remains a challenging problem. These challenges stem from the difficulty of feature engineering and knowledge distillation from the raw sensor data, as well as the computational and memory constraints of battery-operated edge devices. To this end, we propose a framework called MHDeep that utilizes commercially available WMSs and efficient DNN models to diagnose three important mental health disorders: schizoaffective, major depressive, and bipolar. MHDeep uses eight different categories of data obtained from sensors integrated in a smartwatch and smartphone. These categories include various physiological signals and additional information on motion patterns and environmental variables related to the wearer. MHDeep eliminates the need for manual feature engineering by directly operating on the data streams obtained from participants. Because the amount of data is limited, MHDeep uses a synthetic data generation module to augment real data with synthetic data drawn from the same probability distribution. We use the synthetic dataset to pre-train the weights of the DNN models, thus imposing a prior on the weights. We use a grow-and-prune DNN synthesis approach to learn both architecture and weights during the training process. We use three different data partitions to evaluate the MHDeep models trained with data collected from 74 individuals. We conduct two types of evaluations: at the data instance level and at the patient level. MHDeep achieves an average test accuracy, across the three data partitions, of 90.4%, 87.3%, and 82.4%, respectively, for classifications between healthy and schizoaffective disorder instances, healthy and major depressive disorder instances, and healthy and bipolar disorder instances. At the patient level, MHDeep DNN models achieve an accuracy of 100%, 100%, and 90.0% for the three mental health disorders, respectively, based on inference that uses 40, 16, and 22 minutes of sensor data collection from each patient.

Additional Key Words and Phrases: disease diagnosis, health monitoring, mental health disorders, neural networks, synthetic data generation, wearable medical sensors.

ACM Reference Format:
Shayan Hassantabar, Joe Zhang, Hongxu Yin, and Niraj K. Jha. 2022. MHDeep: Mental Health Disorder Detection System based on Wearable Sensors and Artificial Neural Networks. 1, 1 (March 2022), 22 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Author's address: Shayan Hassantabar, Joe Zhang, Hongxu Yin, and Niraj K. Jha, emails:{seyedh,zhaoz,hongxuy,jha}@princeton.edu, Princeton University, Princeton, New Jersey, 08544, USA.

XXXX-XXXX/2022/3-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Mental health problems impact around 20% of the world population [50]. They may negatively affect a person's mind, emotions, behavior, and even physical health. Mental health issues may include various disorders like bipolar, depression, schizophrenia, and attention-deficit hyperactivity, to name but a few. These disorders not only affect adults but children and adolescents may suffer from them as well [43]. Moreover, patients with serious mental health issues have a higher risk of morbidity due to physical health problems.

In order to understand the mental health condition of the patient and provide suitable patient care, early detection is essential. However, this remains a public health challenge. While many other diseases can be diagnosed based on specific medical tests and laboratory measurements, detection of mental health problems mainly relies on self-reports and responses to specific questionnaires designed for identifying certain patterns of behavior and social interactions. Hence, to address this challenge, novel detection strategies are needed.

There has been recent interest in employing machine learning to detect mental health conditions [16]. Neural networks (NNs) are popular machine learning models that use nonlinear computations to make inferences from large datasets. Thus, they have started being deployed in the smart healthcare domain [4, 23, 25, 36, 60, 61].

In previous studies, two main data sources for deep learning-based analysis of mental health have been clinical data and social media usage data. The former includes studies that use neuro-image data for detecting various mental health disorders [48], electroencephalogram (EEG) data to study brain disorders [3], and analysis of electronic health records (EHR) to study mental health problems [19]. Moreover, social media usage patterns have been used to predict the personal traits of the user. As a result, several recent works focus on exploiting such patterns to detect psychiatric illness [7].

Although the above works have demonstrated the promise of using machine learning in identifying mental health disorders, daily mental health monitoring is still a challenge. Because mental health condition treatment delays may lead to negative outcomes, potentially even loss of life, it is desirable to have immediate and pervasive mental health detection. This is the motivation behind our mental health detection system, MHDeep. As shown in Fig. 1, MHDeep relies on physiological data collected using various WMSs. WMSs can be used to continuously monitor the physiological signals of the wearer throughout the day. This enables constant tracking of the health conditions of the user. MHDeep uses various sensors embedded in smartwatches and smartphones. For training purposes, the collected physiological data are processed to obtain a comprehensive dataset. MHDeep combines data from WMSs with the inference capabilities of DNNs to directly extract mental health condition from the physiological signals. These inferences can be communicated to a health server that is accessible to the physician. This has the potential to enhance the ability of the physician to intervene quickly when mental health conditions deteriorate.

The difficulty of data collection and labeling limits the amount of available data. Hence, reducing the cost of this process is of great importance [17]. To this end, MHDeep uses synthetic data drawn from the same probability distribution as real data to augment the dataset. It also leverages a grow-and-prune DNN synthesis approach [14, 27] to train accurate and computationally efficient DNN models to detect the mental health condition of the user.

The major contributions of this article are summarized next.

- We demonstrate an easy-to-use, accurate, and pervasive mental health disorder detection system, called MHDeep. MHDeep combines physiological signals collected from WMSs with the prediction power of DNNs to detect three main mental health disorders: bipolar, major depressive, and schizoaffective. Unlike many other approaches for detecting mental health problems, MHDeep does not rely on any self-reports from the user.
- We do an extensive search to extract the most appropriate set of data categories for detecting each of the three mental health disorders.

- MHdeep relies on a synthetic data generation module to alleviate the concerns arising from the unavailability of large datasets. It uses a grow-and-prune DNN synthesis approach to improve the accuracy of the DNN models while reducing their computational costs.
- We demonstrate the performance, accuracy, and feasibility of MHDeep through extensive evaluations.

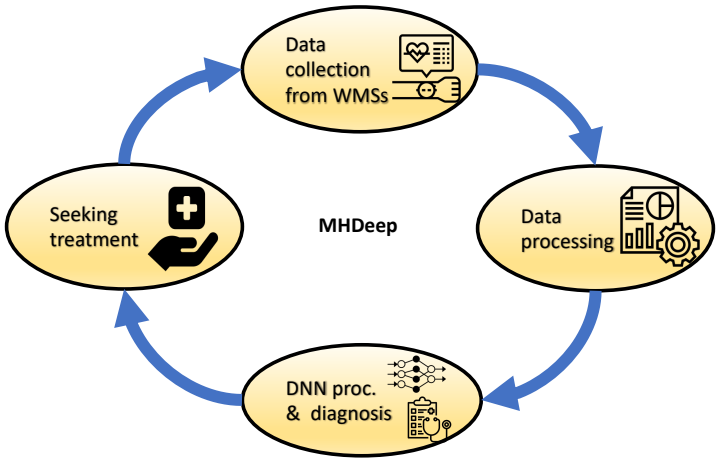

Fig. 1. The MHDeep mental health disorder detection system

The rest of the article is organized as follows. Section 2 presents background information on various works related to MHDeep. Section 3 explains the MHDeep framework thoroughly. Section 4 provides implementation details. Section 5 presents experimental evaluations. Section 6 provides a short discussion related to this work. Finally, Section 7 concludes the article.

2 BACKGROUND

In this section, we first provide background information on various mental health disorders and how they affect patient lives. Next, we discuss various methods for identifying mental health conditions based on machine learning. We also discuss some of the related work on synthesizing efficient DNN models. Finally, we discuss WMSs and their applications to various disease diagnosis frameworks.

2.1 Mental health disorders and their impact

Mental health conditions can affect a patient's thinking, feeling, and behavior. They may have a deep impact on the daily life of the person and affect their ability to adequately perform in society. There are hundreds of different mental illnesses [31]. We discuss the three mental health disorders that we target in this work: bipolar, major depressive, and schizoaffective. These are among the major mental health illnesses and have also been the subject of other related studies [37].

Bipolar disorder can cause a dramatic shift in a person's mood, energy, and behavior. It is characterized by experiences of alternating episodes of manic and depressed states. Major depressive disorder may present different symptoms like loss of interest, sleep disturbance, change in appetite, and feeling of fatigue. Schizoaffective disorder is characterized by various symptoms of schizophrenia such as episodes of hallucinations and delusions. It may also present other symptoms such as disorganized thinking, depressed mood, and manic behavior.

Apart from various conditions that these mental health disorders may cause, stereotypes related to mental health seem to still be widely prevalent in society, not just among uninformed people but even among well-trained professionals [12]. These stereotypes often lead to social and employment discrimination [8] and poor treatment of physical health problems [13].

2.2 Deep learning for mental health

Deep learning has been recently used to better understand and detect mental health problems. Deep learning approaches have been applied to various types of data: mainly clinical data and social media usage data [51]. The three types of clinical data used in these works are neuroimage data, EEG data, and EHR data.

Several studies demonstrate the effectiveness of neuroimages in detecting neuropsychiatric disorders [48]. Two types of neuroimage data used in such works are functional magnetic resonance imaging (fMRI) and structural magnetic resonance imaging (sMRI). fMRI measures brain activity by monitoring blood oxygenation and flow in response to neural activity. sMRI examines the anatomy and pathology of the brain. Deep belief networks have been used to detect the presence of attention-deficit hyperactivity disorder (ADHD) using fMRI and sMRI data [29, 30]. These data types have also been used to detect schizophrenia [42, 62]. Depression has been detected using time-series fMRI data using convolutional neural networks (CNNs) and autoencoders [18]. EEG is another source of data for studying brain disorders. For example, CNN-based feature extraction from EEG data has been used to detect depression [3]. EHR is a collection of patient-centered records and includes both structural data such as laboratory reports and unstructured data such as clinical and discharge notes. Because EHR is a collection of longitudinal records, recurrent neural networks (RNNs) have been used to distill information from them. Pham et al. [41] use RNN architectures to predict future outcomes of depressive episodes. Unstructured clinical notes have also been analyzed with deep learning-based models to detect depression [19]. Social media usage data have also proved their usefulness in identifying psychiatric illnesses. Birnbaum et al. [7] investigate Facebook messages and patterns of sharing images on social media to distinguish among healthy individuals, individuals with a schizophrenia spectrum disorder, and individuals with mood disorders. Other works have used DNN models with textual data and image data shared on social media platforms to detect stress [32], depression [45], and risk of suicide [11].

MHDeep relies on generating the synthetic data from the same distribution as the real data. However, there are several works that develop compact and accurate dynamical models to capture the characteristics of the system under study. The synthetic data can be generated based on these mathematical models. Xue et al. [58] proposed a multivariate fractal model to capture the long-range memory and spatial dependencies that exist in biological processes, and showed the benefits of this approach within the context of the brain-machine-body interface. The authors of [56] proposed a mathematical strategy for constructing models for complex non-linear dynamics. In addition, Xue et al. [57] proposed a polynomial algorithm to obtain a sub-optimal solution with optimality guarantees for the NP-hard problem of determining the minimum number of sensors to enable the recovery of global data dynamics. Although MHDeep uses deep learning for mental health disorder diagnosis, it is worth mentioning that deep learning has also been used in other use cases such as in vaccine discovery for SARS-CoV-2 [59].

2.3 Efficient neural network synthesis

Next, we summarize the main synthesis approaches for obtaining compact DNN models. Conventional synthesis methods are based on the use of efficient building blocks. For example, MobileNetV2 [46] leverages inverted residual blocks to reduce model size and computations significantly. Wu et al. [55] use shift-based operations rather than convolution layers to significantly reduce computational costs of the model. The main drawback of such approaches is the need for considerable design insight and trial-and-error process to design such efficient

building blocks. Network compression is another approach for the design of efficient models. It removes the need for design insights. Network pruning is a widely used method that eliminates weights or filters that do not enhance model performance. Han et al. [22] have shown the effectiveness of pruning in removing redundancy in CNNs and multilayer-perceptron architectures. Grow-and-prune DNN synthesis uses network growth followed by network pruning in an iterative process to improve model performance while ensuring its compactness [14, 27].

Another recent approach relies on the use of reinforcement learning (RL) to search for DNN architectures in an automated flow. It is known as neural architecture search (NAS) [63]. NAS generally uses a controller, e.g., an RNN, to iteratively generate candidate architectures in the search process. The RL controller is improved based on candidate performance. As an example, MnasNet [52] uses an RL-based approach to develop efficient DNNs for mobile platforms. However, the downside of the RL-based NAS approach is that it is computationally intensive. FBNet [54] uses the Gumbel softmax function to optimize weights and connections using a single objective function. NEAT [49] uses evolutionary algorithms to generate optimized and increasingly complex architectures over multiple generations. Combining efficient evolutionary search algorithms with various performance predictors, e.g., for accuracy, energy, and latency, is another approach for synthesizing accurate yet compact CNNs and DNNs [15, 24].

2.4 Wearable medical sensors

Due to recent developments in low-power sensor design and efficient wireless communication, battery-powered WMSs are becoming ubiquitous. More than 123 million WMSs were sold worldwide in 2018. This number is projected to grow to 1 billion by the end of 2022 [2]. WMSs can track different aspects of human health including heart rate, body/skin temperature, respiration rate, blood pressure, EEG, electrocardiogram (ECG), and Galvanic skin response (GSR) [5]. Furthermore, the number of physiological signals that can be measured using WMSs keeps growing every year.

WMSs have begun to be used in many smart healthcare applications. CodeBlue [33] is a sensor network that collects vital health signs and transmits them to the healthcare provider. MobiHealth [53] is a WMS-based body-area network (BAN) that realizes an end-to-end mobile health monitoring platform. Yin et al. [61] use WMSs for pervasive diagnosis of Type-I and Type-II diabetes. CovidDeep [25] is a WMS-based framework for quick detection of SARS-CoV-2/COVID-19.

For data collection in MHDeep, we use an Empatica E4 smartwatch [1] to record a subset of patient's physiological signals. It is a wearable wireless device designed for comfortable, continuous, and real-time data acquisition. We also use a smartphone to simultaneously record signals related to motion information and environmental variables. Because the DNNs developed for diagnosing various mental health conditions can reside on the smartphone, use of a smartwatch/smartphone-based BAN can enable accurate, yet convenient, disease diagnosis and continuous healthcare monitoring.

3 METHODOLOGY

In this section we describe various parts of the MHDeep framework. First, we give an overview of our approach. Then, we discuss the data collection and preparation process, synthetic data generation, and grow-and-prune DNN synthesis.

3.1 The MHDeep framework

We illustrate the MHDeep framework in Fig. 2. The input data are derived from the physiological signals collected using various WMSs in the smartwatch and smartphone in a non-invasive, passive, and efficient manner. The list of collected data streams include GSR, skin temperature (ST), inter-beat interval (IBI), and

3-way acceleration (tri-axial accelerometer) from the smartwatch. In addition, some information related to the motion patterns of the user and ambient information are collected using smartphone sensors. This includes ambient temperature, gravity, acceleration, and angular velocity. After sensor data collection, the collected signals are synchronized, aggregated, and merged into a comprehensive data input for subsequent analysis. To enhance the accuracy of subsequent analysis and improve noise tolerance, we normalize the data. The process of data collection and preparation is discussed in more detail in Section 3.2. When the size of the training dataset is small, it can be useful to generate a synthetic dataset from the same probability distribution as the real training dataset. MHDeep leverages Gaussian mixture model (GMM)-based density estimation to generate the synthetic data. Then, it uses grow-and-prune DNN synthesis to generate inference models that are both accurate and computationally efficient. Section 3.3 discusses the MHDeep DNN synthesis process in detail. MHDeep generates DNN architectures that are efficient enough to be deployed on the edge devices such as smartphones or smartwatches. Section 3.4 discusses the inference process of the MHDeep DNN models for diagnosis and daily monitoring of mental health issues.

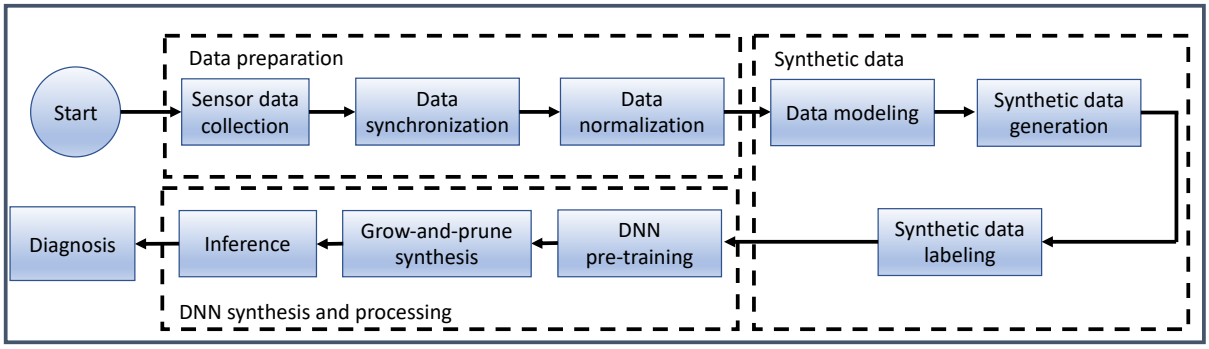

Fig. 2. Schematic diagram of the MHDeep framework.

3.2 Data collection and preparation

We collected WMS data from a total of 74 adult participants at the Hackensack Meridian Health Carrier Clinic, Belle Mead, New Jersey. The participants were diagnosed by medical professionals at the clinic. The 74 participants comprised the following four categories: 25 healthy participants (no mental health disorder), 23 participants with bipolar disorder, 10 participants with major depressive disorder, and 16 participants with schizoaffective disorder. The experimental procedure for data collection and analysis was approved by the Institutional Review Board of Princeton University. The physiological signals of the participants were captured by a commercially-available Empatica E4 smartwatch [1] and a Samsung Galaxy S4 smartphone, as shown in Fig. 3. First, we collected data from an extensive range of sensors embedded in both the smartwatch and smartphone. We analyzed the mean value and standard deviation of collected data from each sensor for different cohorts. We identified a final set of eight sensors as being the most informative in terms of distinguishing between these four cohorts. Table 1 summarizes the final set of data types that we used in this study. The physiological signals are derived from WMSs embedded in the smartwatch. They include GSR that measures sympathetic nervous system arousal, IBI that indicates the heart rate, ST that provides skin temperature readings, and 3-axis accelerometer (Acc-W) that measures acceleration in the x, y, and z directions. The information collected from these sensors is useful for detecting various mental disorders. For example, the electrodermal response can be used as a feature to detect patients affected by depression disorder or to detect different mood disorders [47]. Bipolar

disorder is associated with cardiac autonomic dysregulation [9] that has an impact on IBI. Skin temperature can be used as a feature to detect stress [28] and bipolar [39] disorders. In addition to the physiological signals, ambient and motion information is also captured using sensors in the smartphone. These include ambient temperature (Temp), gravity (Grav), acceleration (Acc-P), and angular velocity (Vel). The motion and ambient information may also be informative in detecting the mental state of the user. For example, Berle et al. [6] show that motor activities of schizophrenic and depressed patients are significantly reduced. In addition to the motion information, ambient temperature can also impact the mental state of the individual and affect the severity of the mental disorder. Mullins et al. [38] argue that cold temperatures can reduce the adverse effect of mental disorders, whereas hot temperatures can exacerbate them. In addition, it is worth mentioning that the acceleration sensors in the smartphone and smartwatch have different sampling rates, and capture different motion information.

Before data collection, all participants are informed about the experiment and are asked to sign a consent form. The data collection setup consists of placing the Empatica E4 smartwatch on the wrist of the participant's non-dominant hand and placing the Samsung Galaxy S4 smartphone in the opposite front pocket. For all participants, we maintain the same orientation for the phone. Data collection lasts around 1.5 hours, during which time the participant is allowed to freely move around in the room with their on-body devices. During this time, the smartwatch and smartphone continuously record and store physiological signals and ambient/motion information. At the end of the data collection period, we remove the smartwatch from the patient's wrist and the smartphone from their pocket. We use the Empatica E4 Connect portal for smartwatch data retrieval. We use a private Android application to download the smartphone data streams. All of the recorded data are timestamped at the time of sampling.

Next, we preprocess the dataset for use in DNN training. We first synchronize the smartwatch and smartphone data streams for each participant. This is necessary because the WMS data streams may vary in their start times and frequencies. Then, we divide the data for each participant into 15-second windows. This window size was chosen based on experiments with the validation set, as discussed later. Each 15-second window of the combined smartwatch/smartphone data constitutes one data instance. There is no time overlap between data instances. To obtain each data instance, we flatten and concatenate the data within the same time window from both the smartwatch and smartphone. This results in a feature space of dimension 2325. The smartwatch (smartphone) contributes 1575 (750) features. All the smartphone sensors have a sampling rate of 5Hz. In addition, the smartwatch sensors include one data stream at 32Hz, two data streams at 4Hz, and one data stream at 1Hz.

Because the participants are in a room during the data collection process, they do not enjoy a wide range of motions, and hence the higher sampling rates for the smartphone sensors are not needed. In addition, the Empatica E4 used for data collection is a medical-grade smartwatch that is designed to capture various physiological signals with their optimal sampling rates. Although collecting data from more sensors at higher sampling rates may provide more information, unnecessarily high sampling rates can lead to a decrease in the battery life of the device. In addition, by targeting a window of 15 seconds for each data instance, we can remedy the low sampling frequency of some of the sensors by considering multiple sensor readings in each data instance.

For each classification task, because the number of individuals in each of the four categories (healthy and three disorders) is small, we created three different data partitions for evaluation. We used circular shifts on a list of numbers denoting patients in each category to create these three data partitions. The value of the stride used for the circular shift is equal to the number of test individuals in each group (as explained later). The data instances extracted from the individuals in each of the four groups (healthy, schizoaffective, depressive, and bipolar) were divided into three sets: training, validation, and test. To evaluate the models on different unseen patients, data instances included in the training, validation, and test sets came from different individuals, i.e.,

Fig. 3. An Empatica E4 smartwatch (left) and Samsung Galaxy S4 smartphone (right) used in the data collection process.

Table 1. Data types collected in the MHDeep framework

Data type	Sampling rate (Hz)	Data source
Galvanic skin response (μS)	4	Smartwatch
Skin temperature ($^{\circ}C$)	4	Smartwatch
Inter-beat interval (ms)	1	Smartwatch
Acceleration (x, y, z)	32	Smartwatch
Ambient temperature ($^{\circ}C$)	5	Smartphone
Gravity (x, y, z)	5	Smartphone
Acceleration (x, y, z)	5	Smartphone
Angular velocity (x, y, z)	5	Smartphone

no individual contributed data to more than one of these sets. Among the healthy participants, for each of the three data partitions, data instances from 15 individuals (60% of the healthy participants) are selected for the training set, from 5 individuals (20% of the healthy participants) for the validation set, and from the remaining 5 individuals (20% of the healthy participants) for the test set. For individuals with bipolar disorder, the training, validation, and test sets contain data instances from 13, 5, and 5 participants, respectively. Among the participants who had major depressive disorder, data instances from 6 participants are selected for the training set and from 2 participants each for the validation and test sets. For individuals with schizoaffective disorder, the training, validation, and test sets include data instances from 10, 3, and 3 participants, respectively.

We create the final dataset for each binary classification task (healthy vs. the mental health disorder) by combining the training, validation, and test sets of the two classes involved in that task. We use SMOTE [10] to up-sample data instances from the minority class. The up-sampling is applied only to the training set. Table 2 shows the number of instances for each classification task for all three data partitions.

Table 2. Details of various datasets (MDD: major depressive disorder).

Classification task	Data partition	Training instances (#individuals)	Validation instances (#individuals)	Test set (#individuals)
Healthy vs. Bipolar	1	14828 (28)	3582 (10)	3754 (10)
Healthy vs. MDD	1	14828 (21)	2414 (7)	2515 (7)
Healthy vs. Schizo.	1	14828 (25)	2789 (8)	3047 (8)
Healthy vs. Bipolar	2	13330 (28)	3922 (10)	3582 (10)
Healthy vs. MDD	2	13330 (21)	3266 (7)	2414 (7)
Healthy vs. Schizo.	2	13330 (25)	3773 (8)	2789 (8)
Healthy vs. Bipolar	3	12054 (28)	4102 (10)	3922 (10)
Healthy vs. MDD	3	12054 (21)	3088 (7)	3266 (7)
Healthy vs. Schizo.	3	12054 (25)	3522 (8)	3773 (8)

3.3 MHDeep DNN synthesis

Fig. 4 shows the DNN architectures used in the MHDeep framework. The architectures receive the input data at the bottom and make their diagnostic decisions at the top. For the healthy vs. major depressive disorder and healthy vs. schizoaffective disorder binary classification tasks, the DNN architecture has four layers with a width of 256, 128, 128, and 2, respectively. For the healthy vs. bipolar disorder binary classification task, we use an DNN architecture with five layers with a width of 256, 128, 64, 32, and 2, respectively. We selected these architectures by verifying the performance of various DNNs (with different numbers of layers and number of neurons per layer) on the validation set and picking the best-performing one. These architectures are initially fully-connected.

We then apply three sequential steps: (i) synthetic data generation to mimic the distribution of the real training data, (ii) pre-training of the DNN architectures with the synthetic data, and (iii) grow-and-prune DNN synthesis to reduce the redundancy of the model while improving its performance. Next, we discuss each step in more detail.

(1) **Synthetic data generation**: In this step, we generate a synthetic dataset that mimics the probability distribution of the real training dataset. Fig. 5 illustrates the synthetic data generation process. We utilize the approach proposed in [26] in order to alleviate the need for large datasets to train DNN architectures. We first use GMM to estimate the density of the training dataset. We optimize the number of mixtures in the GMM by monitoring the likelihood of validation data. We choose the number of components that maximizes the following criterion:

$$N^* = \underset{N}{arg\,max}(\texttt{GMM}_N(x).\texttt{score}(X_{validation}))$$

Finally, we train the optimal GMM model with N^* mixtures on the combination of the training and validation datasets. By sampling this model, we are able to generate the synthetic data:

$$X^* = GMM_{N^*}(X_{total}).sample()$$

For our experiments, we generate 100,000 synthetic data instances. The final step is labeling of the synthetic dataset. We use a traditional machine learning model for this purpose. We evaluate various models, e.g., the support vector machine and random forest models based on different splitting criteria (such as Gini index and entropy), and different depth limits on the decision trees, on the validation set.

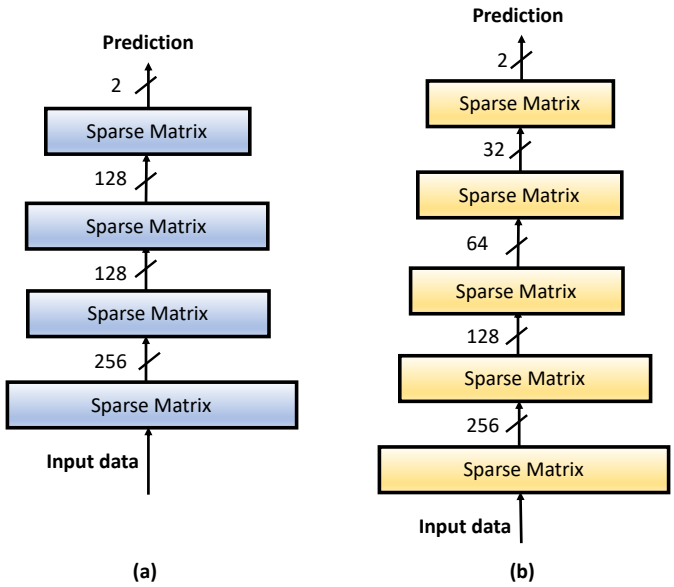

Fig. 4. Architecture of MHDeep DNNs for: (a) healthy vs. major depressive disorder and healthy vs. schizoaffective disorder, and (b) healthy vs. bipolar disorder.

The model with the highest accuracy is used to label the synthetic data. Note that because synthetic data are only used to pre-train an DNN (with subsequent training with real data), the accuracy of the support vector machine or random forest model is not a critical factor.

(2) **DNN pre-training**: In this step, we use the labeled synthetic data to obtain a prior on the weights of the DNN architecture by pre-training them. The intuition behind this step is that pre-training the DNN provides a suitable inductive bias to the parameters of the DNN. As a result, we can commence with the final training stage with a better weight initialization. Therefore, it alleviates the need for large training datasets. By using this methodology, we obtain models that are more accurate than both the traditional machine learning model used for labeling and the DNN model trained only on the real dataset.

(3) **Grow-and-prune DNN synthesis**: MHDeep uses a grow-and-prune DNN synthesis paradigm to train the models. Algorithm 1 summarizes this process. It uses a mask-based approach. For each weight matrix, there is an associated binary mask of the same size that is used to disregard dormant connections in the architecture. It applies magnitude-based pruning and full growth to fully-connected DNNs iteratively. For magnitude-based pruning, a hyperparameter α is used to depict the pruning ratio. We prune a connection if and only if its weight is in the lowest $\alpha * 100$ percent of the weights in its associated layer. Finally, for the pruned connections, we set the weight and its binary mask both to 0. Because connection pruning is an iterative process, we retrain the network to recover its performance after each pruning iteration. We then grow the network to restore all its connections. In our experiments, after each architecture-changing operation, we train the DNN for 20 epochs. In addition, we set the number of iterations to 5. We evaluate the model on the validation set after each epoch, and we record the learned weights and masks of the pruned model with the highest validation accuracy.

Algorithm 1 Grow-and-prune synthesis

Input: Pre-trained DNN architecture; iteration number $numIterations$; weight matrix $W \in R^{M \times N}$; mask matrix $Mask$ of the same dimension as the weight matrix; α: pruning ratio

```
best-validation-acc = 0
for all layers in the DNN do
    t = (α × MN)^th largest element in |W|
    for all w_ij do
        if |w_ij| < t then
            Mask_ij = 0
        end if
    end for
    W = W ⊗ Mask
end for
Train DNN for given #epochs
validation-acc = evaluate DNN on validation set
if validation-acc > best-validation-acc then
    best-validation-acc = validation-acc
    Save the DNN
end if
for numIterations do
    for all layers in the DNN do
        Mask_[1:M,1:N] = 1
    end for
    Train DNN for given #epochs
    for all layers in the DNN do
        t = (α × MN)^th largest element in |W|
        for all w_ij do
            if |w_ij| < t then
                Mask_ij = 0
            end if
        end for
        W = W ⊗ Mask
    end for
    Train DNN for given #epochs
    validation-acc = evaluate DNN on validation set
    if validation-acc > best-validation-acc then
        best-validation-acc = validation-acc
        Save the DNN
    end if
end for
```

Output: Best architecture with the weight and mask matrices

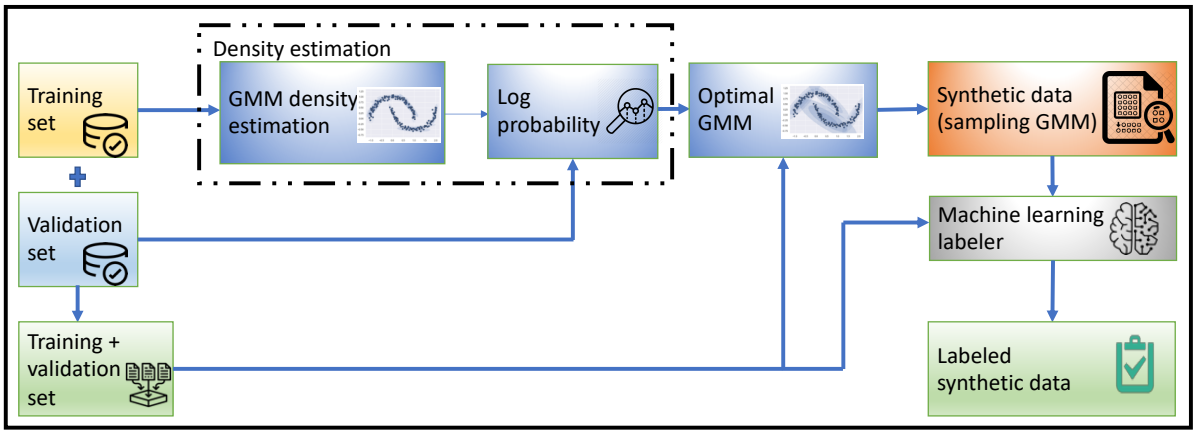

Fig. 5. Schematic diagram of the MHDeep synthetic data generation process.

3.4 MHDeep inference process

The trained DNN models can be used for diagnosis or daily monitoring of the mental state of the user based on a collection of physiological signals and ambient information during the day. The collected data streams are processed based on the step explained in Section 3.2. We feed the processed data to the MHDeep DNN models that predict the mental health condition of the user. This information can then be sent to the physician for early treatment.

4 IMPLEMENTATION DETAILS

In this section, we give an overview of the hardware and software packages we used in the implementation of the MHDeep framework. We have implemented the data processing and preparation parts of the MHDeep framework in Python and the MHDeep DNN synthesis framework in PyTorch. We use the Nvidia Tesla P100 data center accelerator for DNN training and evaluation. We use the cuDNN library to accelerate GPU processing. For training, we use a stochastic gradient descent (SGD) optimizer, with a learning rate of 5e-4 and a batch size of 256. We use 100,000 synthetic data instances to pre-train the network architecture. In the grow-and-prune synthesis phase, we train the network for 20 epochs each time the architecture changes. We use an SGD optimizer, with an initialized learning rate of 1e-4 that we halve in each succeeding iteration. We apply network-changing operations over five iterations.

5 EVALUATION

In this section, we analyze the performance of MHDeep DNN models for diagnosing three mental health disorders. This entails three binary classifications: (i) schizoaffective disorder vs. healthy individuals, (ii) major depressive disorder vs. healthy individuals, and (iii) bipolar disorder vs. healthy individuals. For each classification task, we use three different data partitions, each partition with data instances obtained from different individuals in the training, validation, and test sets.

The MHDeep DNN models are evaluated with four different metrics: test accuracy, false positive rate (FPR), false negative rate (FNR), and F1 score. Accuracy measures overall classification performance. It is simply the ratio of all the correct predictions on the test data instances and the total number of such instances. FPR and FNR measure how often healthy individuals are declared to have the corresponding mental health condition

and *vice versa*, respectively. In addition to these four metrics, we also report the TPR (sensitivity) and the TNR (specificity) values that are equal to 1 minus the FNR and FPR values, respectively.

We conduct two different performance evaluations: at the data instance level and the patient level. First, we report the performance of the MHDeep DNN models in detecting each of the three mental health disorders at the data instance level. Next, we evaluate the accuracy of the models in detecting mental health disorders at the patient level.

5.1 MHDeep performance evaluation at the data instance level

We first analyze the performance of the three binary classifiers. We begin by training DNN models on features obtained from subsets of the eight data categories presented in Table 1. We analyze all the subsets of the eight data categories and report results for the top models. Because there are eight data categories, there are 256 subsets, with one being the null subset. We evaluated the remaining 255 subsets. This helps distinguish the impact of each data category and to find the most effective combination of categories for each classification task. Next, we highlight the best-performing data categories for each of the three classification tasks. We then compare the performance of MHDeep DNNs with that of traditional machine learning models. Furthermore, we conduct an ablation study that shows the impact of each step of MHDeep DNN training.

Table 3 shows the results of classification between healthy and schizoaffective data instances. The best data category subset, in this case, achieves an average test accuracy of 90.4%. We also report test accuracy, FPR, FNR, TPR, TNR, and F1 score for each of the three data partitions. The model reaches the highest test accuracy of 93.3% on the second data partition. Furthermore, for the healthy instances, the top model achieves a low average FPR of 6.5%, demonstrating its effectiveness in avoiding false alarms. For the schizoaffective instances, the model achieves an average FNR of 16.9%, indicating reasonable effectiveness in raising alarms when schizoaffective disorder does occur. We report the number of parameters (#params) and floating-point operations (FLOPs) required for each model. We also compare #params and FLOPs of the models with those of the fully-connected baselines. As can be seen, using the grow-and-prune DNN synthesis approach enables us to reduce both #params and FLOPs, leading to a reduction in memory and computational requirements.

We present the results for classification between healthy and major depressive disorder instances in Table 4. The data category subset with the best performance achieves an average test accuracy of 87.3%. This model achieves the highest accuracy of 91.2% on the second data partition. It achieves an average FPR (FNR) of 6.8% (29.3%).

Table 5 presents the results for classification between healthy and bipolar disorder instances. In this case, the model trained on the best data category subset achieves an average test accuracy of 82.4%, with an FPR (FNR) of 16.7% (20.7%).

As the results for these three classification tasks show, there is some variability in test performance on the three data partitions. Because the number of represented individuals in each classification task is limited and each partition consists of different individuals in the training, validation, and test sets, performance variability on the test set can be expected. Hence, we report average values across the three data partitions for each metric. In addition, by comparing the results of the three classification tasks, we can infer that it is more difficult to detect bipolar disorder than the other two disorders. This may be due to the nature of the bipolar disorder, which may induce various mood changes. As a result, the physiological signals of the patient can change during these changes of moods, leading to a more difficult diagnosis. It is also worth mentioning that the subset of data categories that leads to the best-performing models for each classification task differs from each other. This points to the need for accurate sensor subset selection.

Next, we compare the performance of our methodology with that of traditional machine learning methods. We present the results for the first data partition in Table 6. We report the results for two data categories for each

Table 3. Test accuracy, FPR, FNR, and F1 score (all in %) for top data categories for classification between healthy and schizoaffective disorder data instances

Data category	Data partition	#Params (compression)	FLOPs (compression)	Acc.	FPR (TNR)	FNR (TPR)	F1 Score
(Acc-P, Temp, Vel, Acc-W, GSR, IBI)	1	275.0k (2.1×)	549.5k (2.0×)	91.2	11.4 (88.6)	5.8 (94.2)	89.5
	2	300.0k (1.9×)	599.5k (1.9×)	81.4	6.6 (93.4)	37.8 (62.2)	72.0
	3	275.0k (2.1×)	549.5k (2.0×)	87.3	2.4 (97.6)	34.1 (65.9)	84.2
Average				86.6	6.8 (93.2)	25.9 (74.1)	81.9
(Acc-P, Temp, Vel, Acc-W, GSR)	1	200.0k (2.8×)	399.5k (2.8×)	90.5	13.0 (87.0)	4.4 (95.6)	89.3
	2	300.0k (1.9×)	599.5k (1.8×)	93.3	4.0 (96.0)	13.3 (86.7)	89.8
	3	250.0k (2.3×)	499.5k (2.2×)	87.5	2.5 (97.5)	32.9 (67.1)	77.9
Average				90.4	6.5 (93.5)	16.9 (83.1)	85.7
(Acc-P, Temp, Grav, Vel, Acc-W, GSR, IBI)	1	300.0k (2.1×)	599.5k (2.0×)	88.3	14.4 (85.6)	7.7 (92.3)	86.8
	2	350.0k (1.8×)	699.5k (1.8×)	82.6	5.9 (94.1)	45.7 (54.3)	66.3
	3	300.0k (2.1×)	599.5k (2.0×)	88.7	3.9 (96.1)	26.4 (73.6)	81.1
Average				86.5	8.1 (91.9)	26.6 (73.4)	78.1

Table 4. Test accuracy, FPR, FNR, and F1 score (all in %) for top data categories for classification between healthy and major depressive disorder data instances

Data category	Data partition	#Params (compression)	FLOPs (compression)	Acc.	FPR (TNR)	FNR (TPR)	F1 Score
(Temp, Grav, Vel, GSR)	1	75.0k (2.7×)	149.5k (2.4×)	89.0	4.6 (95.4)	26.7 (73.3)	79.4
	2	120.0k (1.7×)	239.5k (1.5×)	90.5	4.5 (95.5)	21.7 (78.3)	82.7
	3	150.0k (1.3×)	299.5k (1.2×)	81.7	12.7 (87.3)	37.9 (62.1)	60.2
Average				87.1	7.3 (92.7)	28.8 (71.2)	74.1
(Acc-P, Grav, Vel, GSR)	1	120.0k (2.0×)	239.5k (1.8×)	88.2	6.5 (93.5)	24.8 (75.2)	78.7
	2	145.0k (1.6×)	289.5k (1.5×)	91.2	1.9 (98.1)	25.7 (74.3)	83.0
	3	185.0k (1.3×)	369.5k (1.2×)	82.4	11.9 (88.1)	37.5 (62.5)	61.3
Average				87.3	6.8 (93.2)	29.3 (70.7)	74.3

classification task. As we can see, our methodology outperforms traditional machine learning models in all cases. This points to the difficulty of distilling information from raw data using traditional machine learning models and highlights the superiority of our methodology in diagnosing various diseases. It is worth mentioning that we achieved similar results on the second and third data partitions.

We also performed an ablation study to analyze the impact of three training methods on the final performance, including training only on real data, pre-training with synthetic data, and grow-and-prune synthesis. We present

Table 5. Test accuracy, FPR, FNR, and F1 score (all in %) for top data categories for classification between healthy and bipolar disorder data instances

Data category	Data partition	#Params (compression)	FLOPs (compression)	Acc.	FPR (TNR)	FNR (TPR)	F1 Score
(Acc-P, Temp, Grav, Vel, Acc-W, IBI, ST)	1	500.0k (1.2×)	999.5k (1.2×)	76.1	47.1 (52.9)	2.8 (97.2)	81.0
	2	480.0k (1.3×)	959.5k (1.3×)	81.4	1.0 (99.0)	34.3 (65.7)	78.9
	3	400.0k (1.6×)	799.5k (1.5×)	89.8	2.1 (97.9)	25.1 (74.9)	83.8
Average				82.4	16.7 (83.3)	20.7 (79.3)	81.2
(Temp, Grav, Vel, Acc-W, GSR, IBI)	1	490.0k (1.2×)	979.5k (1.1×)	75.6	47.1 (52.9)	3.8 (96.2)	80.5
	2	450.0k (1.3×)	899.5k (1.2×)	81.6	0.9 (99.1)	34.2 (65.8)	79.0
	3	380.0k (1.5×)	759.5k (1.5×)	87.0	2.1 (97.9)	33.0 (67.0)	78.4
Average				81.4	16.7 (83.3)	23.7 (76.3)	79.3

performance results for the first data partition in Table 7. For each classification task, we report the performance of two data categories. We can see that using the synthetic dataset to pre-train the weights of the model helps improve performance in most cases, thanks to a better initialization point for the final training process. In addition, the application of grow-and-prune synthesis yields models specialized for the task at hand and not only results in more compact models (as shown in Tables 3, 4, 5), it improves the performance of the models as well. It is worth mentioning that similar results were obtained on the other two data partitions.

5.2 MHDeep performance evaluation at the patient level

Next, we show patient-level diagnostic test accuracy. We use the most accurate model from among the models discussed above for each classification task. Fig. 6 shows the results. In these graphs, we plot patient-level test accuracy vs. the duration of sensor data collection needed for inference. Prediction is performed for each patient by simply taking the majority of the predicted labels over all data instances in the given sensor data collection duration. As discussed earlier, each data instance is composed of a 15-second window of sensor data. We step up the data duration size by 2 minutes each time. Thus, we add eight data instances in each 2-minute window. By predicting the label for each participant, we define the final test accuracy as the ratio of the participants that are correctly diagnosed over the number of participants in the test set. As we can see, the models reach 100% test accuracy after a certain point for distinguishing healthy individuals from those with schizoaffective and major depressive disorders. In addition, the best model for classification between healthy and bipolar disorder individuals reaches 90.0% patient-level accuracy. Table 8 shows the minimum sensor data collection duration needed to reach saturation accuracy. The durations are 40, 16, and 22 minutes for healthy vs. schizoaffective disorder, healthy vs. major depressive disorder, and healthy vs. bipolar disorder classifications, respectively. Not surprisingly, these results indicate that it is easier to obtain higher test accuracies at the patient level than at the data instance level because the former is simply based on taking the majority of the predictions over all instances in the data duration.

Although the MHDeep DNNs have been evaluated on the same platform we used for training, they can be encapsulated in a smartphone app for diagnosis of mental disorders. Through the MHDeep app, the user can be instructed to wear the Empatica E4 smartwatch and correctly position the smartphone before data collection begins. The MHDeep preprocessing pipeline can be used to normalize the data using the minimum and

Table 6. Comparison with machine learning models on first data partition

Classification Task (Data category)	Method	Accuracy (%)	FPR (%)	FNR (%)	F1 score (%)
Healthy vs. Schizo. (Acc-P, Temp, Vel, Acc-W, GSR, IBI)	SVM	79.0	31.8	5.7	79.0
	Decision tree	68.9	44.0	12.9	68.8
	Random forest	71.9	42.6	7.3	71.9
	k-nearest neighbor ($k = 6$)	72.4	37.9	12.8	72.5
	Naive Bayes	70.4	30.0	29.1	70.0
	AdaBoost	74.9	30.4	17.7	74.7
	Our DNN	91.2	11.4	5.8	89.5
Healthy vs. Schizo. (Acc-P, Temp, Vel, Acc-W, GSR)	SVM	77.2	33.9	7.0	77.2
	Decision tree	68.3	43.1	15.4	68.3
	Random forest	67.1	48.2	11.0	67.0
	k-nearest neighbor ($k = 15$)	75.2	34.4	11.1	75.2
	Naive Bayes	70.0	30.1	29.9	69.9
	AdaBoost	75.5	29.6	17.2	75.4
	Our DNN	90.5	13.0	4.4	89.3
Healthy vs. MDD (Temp, Grav, Vel, GSR)	SVM	74.8	15.7	48.6	68.4
	Decision tree	52.4	49.1	43.9	50.4
	Random forest	68.7	25.5	45.6	63.6
	k-nearest neighbor ($k = 12$)	67.4	30.2	38.4	63.7
	Naive Bayes	67.0	35.8	26.3	64.9
	AdaBoost	56.1	46.3	38.0	54.2
	Our DNN	89.0	4.6	26.7	79.4
Healthy vs. MDD (Acc-P, Grav, Vel, GSR)	SVM	75.5	16.8	43.4	70.0
	Decision tree	52.7	42.5	58.9	48.4
	Random forest	70.0	24.5	43.5	65.1
	k-nearest neighbor ($k = 12$)	62.6	36.2	40.4	59.4
	Naive Bayes	61.2	45.1	23.3	60.1
	AdaBoost	54.4	48.5	38.7	52.7
	Our DNN	88.2	6.5	24.8	78.7
Healthy vs. Bipolar (Acc-P, Temp, Grav, Vel, Acc-W, IBI, ST)	SVM	68.1	47.6	17.6	67.1
	Decision tree	63.8	48.0	25.5	63.1
	Random forest	61.3	54.4	24.3	60.1
	k-nearest neighbor ($k = 10$)	70.9	52.5	7.9	68.8
	Naive Bayes	69.1	40.7	21.9	68.6
	AdaBoost	60.1	59.7	21.9	58.1
	Our DNN	76.1	47.1	2.8	81.0
Healthy vs. Bipolar (Temp, Grav, Vel, Acc-W, GSR, IBI)	SVM	71.4	43.0	15.5	70.5
	Decision tree	64.3	52.3	20.6	63.0
	Random forest	65.8	54.7	15.6	63.9
	k-nearest neighbor ($k = 7$)	64.6	50.1	22.1	63.5
	Naive Bayes	61.0	41.6	36.6	60.9
	AdaBoost	67.1	42.4	24.3	66.6
	Our DNN	75.6	47.1	3.8	80.5

Table 7. Impact of different training methods on the performance of the model

Classification Task (Data category)	Training method	Accuracy (%)	FPR (TNR) (%)	FNR (TPR) (%)
Healthy vs. Schizo. (Acc-P, Temp, Vel, Acc-W, GSR, IBI)	Real training dataset	87.0	18.4 (81.6)	5.2 (94.8)
	Real+synthetic training dataset	89.6	6.1 (93.9)	20.8 (79.2)
	Real+synthetic training dataset + grow-prune	91.2	11.4 (88.6)	5.8 (94.2)
Healthy vs. Schizo. (Acc-P, Temp, Vel, Acc-W, GSR)	Real training dataset	87.8	18.2 (81.8)	3.7 (96.3)
	Real+synthetic training dataset	86.2	19.6 (80.4)	5.5 (94.5)
	Real+synthetic training dataset + grow-prune	90.5	13.0 (87.0)	4.4 (95.6)
Healthy vs. MDD (Temp, Grav, Vel, GSR)	Real training dataset	85.0	7.7 (92.3)	33.1 (66.9)
	Real+synthetic training dataset	85.8	10.4 (89.6)	23.6 (76.4)
	Real+synthetic training dataset + grow-prune	89.0	4.6 (95.4)	26.7 (73.3)
Healthy vs. MDD (Acc-P, Grav, Vel, GSR)	Real training dataset	84.4	11.3 (88.7)	26.0 (74.0)
	Real+synthetic training dataset	85.4	8.2 (91.8)	30.4 (69.6)
	Real+synthetic training dataset + grow-prune	88.2	6.5 (93.5)	24.8 (75.2)
Healthy vs. Bipolar (Acc-P, Temp, Grav, Vel, Acc-W, IBI, ST)	Real training dataset	74.6	48.9 (51.1)	3.9 (96.1)
	Real+synthetic training dataset	75.9	47.4 (52.6)	2.8 (97.2)
	Real+synthetic training dataset + grow-prune	76.1	47.1 (52.9)	2.8 (97.2)
Healthy vs. Bipolar (Temp, Grav, Vel, Acc-W, GSR, IBI)	Real training dataset	74.4	49.9 (50.1)	3.5 (96.5)
	Real+synthetic training dataset	75.3	46.7 (53.3)	4.6 (95.4)
	Real+synthetic training dataset + grow-prune	75.6	47.1 (52.9)	3.8 (96.2)

maximum values used during training and divide the data into 15-second window data instances. Finally, using DNN models for various mental disorders, we can obtain average prediction probabilities. We can diagnose the mental disorder by comparing this average with a user-defined threshold. The MHDeep app would need a limited amount of battery energy. From the study done in [35], we roughly estimate between 0.5 to 1 Watt of battery power for the application. This is based on a range of power estimates of smartphone apps under various categories, such as educational, social media, and entertainment, studied in [35]. As explained earlier, we need 40 minutes of sensor data collection to reach patient-level saturation accuracy for the healthy vs. bipolar disorder diagnosis. As a result, assuming that MHDeep app processing takes 40 minutes in the worst case and smartphone battery works at 3.8V, this translates to 88-175 mAh energy consumption. For a smartphone, such as Samsung Galaxy S8+ with a battery with 3500 mAh capacity, this results in 2.5% to 5.0% battery consumption for the app. In addition, each DNN inference pass takes 0.6 milliseconds in our desktop implementation. As a result, DNN inference for 160 data instances takes 0.1 seconds. In comparison, the sensor data collection takes 40 minutes or 2400 seconds. Even taking into account that DNN inference may take longer on a smartphone, it

Table 8. Minimum sensor data collection duration (in minutes) needed to reach saturation patient-level accuracy (in %) for each classification task

Classification	Time (mins.)	Saturation accuracy (%)
Healthy vs. Schizoaffective disorder	40	100
Healthy vs. Major depressive disorder	16	100
Healthy vs. Bipolar disorder	22	90.0

would be considerably shorter than the duration of data collection. As a result, the energy consumption of the MHDeep app will be dominated by other components of the smartphone such as Bluetooth connectivity and the display.

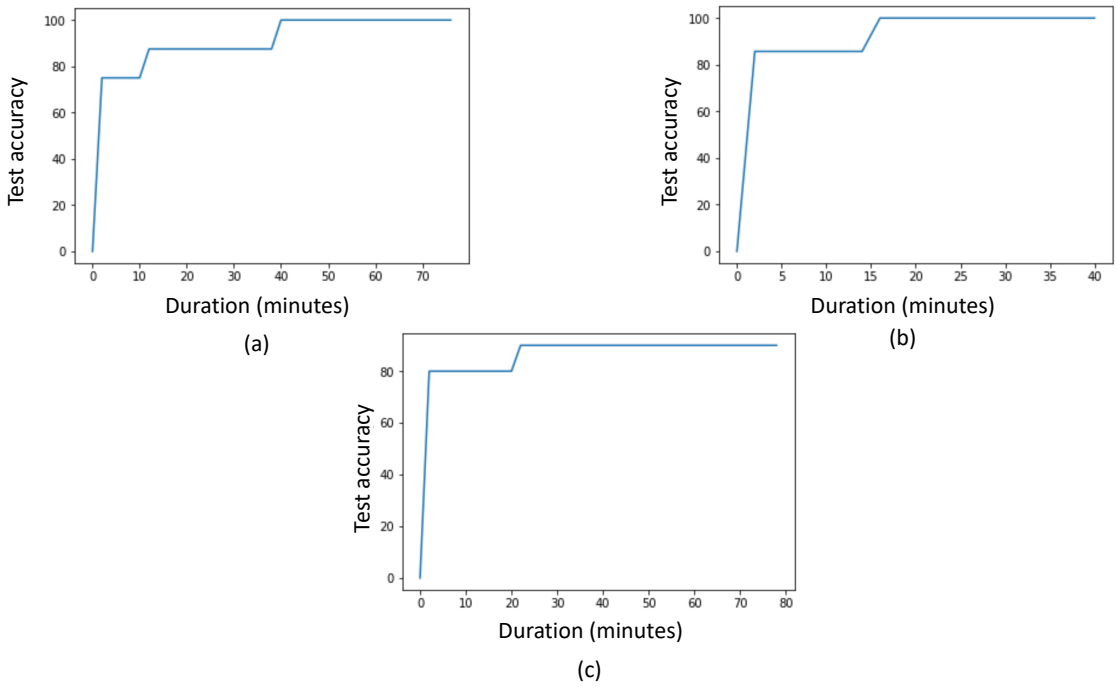

Fig. 6. Patient-level test accuracy vs. sensor data duration needed for classification between (a) healthy and schizoaffective disorder individuals, (b) healthy and major depressive disorder individuals, and (c) healthy and bipolar disorder individuals.

We also report results from some of the related works on use of various machine learning methods for detection of mental health disorders. The results are presented in Table 9. Because these works use different data sources and solve different problems as compared to MHDeep, the goal of this comparison is only to highlight a few related works that target use of machine learning in the mental health domain, the main method, data source, and duration they employ, as well as their main results. As opposed to these methods, MHDeep does not rely on manual feature extraction from various data sources. It directly works on raw sensor data.

Table 9. Comparison with related works

Mental disorder	Method	Data duration	Data sources	Result
Depression	Random forest [44]	Several months	Instagram photos	70% accuracy
Depression	Hypothesis testing[40]	1 week	Accelerometer	Showed reduction in physical activity in depressed patients
MDD	MHDeep	1.5 hours	Physiological, motion, ambient	100% patient-level accuracy
Bipolar	Naive Bayes [21]	12 weeks	Phone call logs, microphone	76% accuracy
	Naive Bayes [20]	12 weeks	GPS, accelerometer	80% accuracy
	Bagging [34]	12 weeks	Accelerometer, microphone, questionnaire	85% accuracy
	MHDeep	1.5 hours	Physiological, motion, ambient	90% patient-level accuracy
Schizophrenia spectrum	Random forest [7]	Several months	Social media posts and messages	0.76 area-under-curve (AUC)
Schizoaffective	MHDeep	1.5 hours	Physiological, motion, ambient	100% patient-level accuracy

6 DISCUSSION

MHDeep combines efficient DNNs with commercially available WMSs to diagnose various mental health disorders. Although several works address mental health problem detection using machine learning, MHDeep is a solution that focuses on an easy-to-use system that can monitor the daily mental health state of the user through their physiological signals, as well as motion and ambient information. The diagnostic decisions can be sent to a health server from where medical professionals can access the information. This can enable them to quickly intervene during severe episodes of the disorder.

Many mental disorders, such as depression, have different stages with different severities. The progress of such mental health problems can impact the patient's life and health in different ways. As a result, it may be useful if the model can predict disease progression over time. Our framework can be extended to predict the progress of mental health disorders by utilizing longitudinal WMS data collected in the training stage. Furthermore, by accumulating more data from each individual, we can synthesize patient-specific models that are specifically designed based on their data. We can obtain such models by fine-tuning the trained general models based on accumulated data from the specific patient.

Diagnosis of mental health disorders is often based on patient's self-report and answers to a questionnaire designed to detect each disorder. In the future, we can improve the performance of MHDeep by adding a specifically designed questionnaire to the data categories. In addition, adding features based on other WMSs such as blood pressure may also help enhance model performance.

7 CONCLUSION

In this article, we proposed a framework called MHDeep that combines data obtained from commercially available WMSs with the knowledge distillation power of DNNs for continuous and pervasive diagnosis of three main mental health disorders: schizoaffective, major depressive, and bipolar. MHDeep uses a synthetic data generation module to address the lack of large datasets. We trained the DNN models by using iterative network growth and pruning to learn both the weights and architecture during the training process. We evaluated MHDeep based on data collected from 74 individuals. It achieves patient-level accuracy of 100%, 100%, and 90.0%, using 40, 16, and 22 minutes of sensor data collected in the inference stage, for classification between healthy and schizoaffective disorder individuals, healthy and major depressive disorder individuals, and healthy and bipolar disorder individuals, respectively. The MHDeep models were also shown to be computationally efficient. Thus, MHDeep can be employed for pervasive diagnosis and daily monitoring while offering high computational efficiency and accuracy.

ACKNOWLEDGMENTS

We would like to thank nurses, physicians, and mental health technicians in Acute Care Unit East and West at the Hackensack Meridian Health Carrier Clinic for smartwatch/smartphone based data collection from various patient cohorts and providing patient labels. Special thanks to Donald J. Parker, CEO of Carrier Clinic, for recognizing the promise of such a study and facilitating data collection, and to Dr. Jacqueline Bienenstock, DNP, RN-BC, Jolene DePaolo, and Lynn Hightower for their help in this project.

REFERENCES

[1] 2020. Empatica E4 connect portal. https://www.empatica.com/connect.
[2] 2020. Number of connected wearable devices worldwide from 2016 to 2022. https://www.statista.com/statistics/487291/global-connected-wearable-devices/
[3] U. Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, Hojjat Adeli, and D. Puthankattil Subha. 2018. Automated EEG-based screening of depression using deep convolutional neural network. *Computer Methods and Programs in Biomedicine* 161 (2018), 103–113.
[4] Ayten Ozge Akmandor and Niraj K. Jha. 2017. Smart health care: An edge-side computing perspective. *IEEE Consumer Electronics Magazine* 7, 1 (2017), 29–37.
[5] Mirza Mansoor Baig and Hamid Gholamhosseini. 2013. Smart health monitoring systems: An overview of design and modeling. *J. Medical Systems* 37, 2 (2013), 9898.
[6] Jan O. Berle, Erik R. Hauge, Ketil J. Oedegaard, Fred Holsten, and Ole B. Fasmer. 2010. Actigraphic registration of motor activity reveals a more structured behavioural pattern in schizophrenia than in major depression. *BMC Research Notes* 3, 1 (2010), 1–7.
[7] Michael L. Birnbaum, Raquel Norel, Anna Van Meter, Asra F. Ali, Elizabeth Arenare, Elif Eyigoz, Carla Agurto, Nicole Germano, John M. Kane, and Guillermo A. Cecchi. 2020. Identifying signals associated with psychiatric illness utilizing language and images posted to Facebook. *Nature Schizophrenia* 6, 1 (2020), 1–10.
[8] James E. Bordieri and David E. Drehmer. 1986. Hiring decisions for disabled workers: Looking at the cause. *J. Applied Social Psychology* 16, 3 (1986), 197–208.
[9] Hsin-An Chang, Chuan-Chia Chang, Nian-Sheng Tzeng, Terry B.J. Kuo, Ru-Band Lu, and San-Yuan Huang. 2014. Heart rate variability in unmedicated patients with bipolar disorder in the manic phase. *Psychiatry and Clinical Neurosciences* 68, 9 (2014), 674–682.
[10] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *J. Artificial Intelligence Research* 16 (2002), 321–357.
[11] Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights* 10 (2018), 1–11.
[12] Patrick W. Corrigan. 2000. Mental health stigma as social attribution: Implications for research methods and attitude change. *Clinical Psychology: Science and Practice* 7, 1 (2000), 48–67.
[13] Patrick W. Corrigan, Dinesh Mittal, Christina M. Reaves, Tiffany F. Haynes, Xiaotong Han, Scott Morris, and Greer Sullivan. 2014. Mental health stigma and primary health care decisions. *Psychiatry Research* 218, 1-2 (2014), 35–38.
[14] Xiaoliang Dai, Hongxu Yin, and Niraj K. Jha. 2019. NeST: A neural network synthesis tool based on a grow-and-prune paradigm. *IEEE Trans. Computers* 68, 10 (2019), 1487–1497.

[15] Xiaoliang Dai, Peizhao Zhang, Bichen Wu, Hongxu Yin, Fei Sun, Yanghan Wang, Marat Dukhan, Yunqing Hu, Yiming Wu, Yangqing Jia, Peter Vajda, Matt Uyttendaele, and Niraj K. Jha. 2019. ChamNet: Towards efficient network design through platform-aware model adaptation. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.

[16] Dominic B. Dwyer, Peter Falkai, and Nikolaos Koutsouleris. 2018. Machine learning approaches for clinical psychology and psychiatry. *Annual Review Clinical Psychology* 14 (2018), 91–118.

[17] Saman Enayati, Ziyu Yang, Benjamin Lu, and Slobodan Vucetic. 2021. A visualization approach for rapid labeling of clinical notes for smoking status extraction. In *Proc. Data Science with Human in the Loop: Language Advances*. 24–30.

[18] Xiang-Fei Geng and Jun-Hai Xu. 2017. Application of autoencoder in depression diagnosis. In *Proc. Int. Conf. Computer Science and Mechanical Automation*.

[19] Joseph Geraci, Pamela Wilansky, Vincenzo de Luca, Anvesh Roy, James L Kennedy, and John Strauss. 2017. Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression. *Evidence-based Mental Health* 20, 3 (2017), 83–87.

[20] Agnes Gruenerbl, Venet Osmani, Gernot Bahle, Jose C. Carrasco, Stefan Oehler, Oscar Mayora, Christian Haring, and Paul Lukowicz. 2014. Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients. In *Proc. Augmented Human Int. Conf.* 1–8.

[21] Agnes Grünerbl, Amir Muaremi, Venet Osmani, Gernot Bahle, Stefan Oehler, Gerhard Tröster, Oscar Mayora, Christian Haring, and Paul Lukowicz. 2014. Smartphone-based recognition of states and state changes in bipolar disorder patients. *IEEE J. Biomedical and Health Informatics* 19, 1 (2014), 140–148.

[22] Song Han, Huizi Mao, and William J. Dally. 2016. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman Coding. In *Proc. Int. Conf. Learning Representations*.

[23] Shayan Hassantabar, Mohsen Ahmadi, and Abbas Sharifi. 2020. Diagnosis and detection of infected tissue of COVID-19 patients based on lung X-Ray image using convolutional neural network approaches. *Chaos, Solitons & Fractals* (2020), 110170.

[24] Shayan Hassantabar, Xiaoliang Dai, and Niraj K. Jha. 2019. STEERAGE: Synthesis of neural networks using architecture search and grow-and-prune methods. *arXiv preprint arXiv:1912.05831* (2019).

[25] Shayan Hassantabar, Novati Stefano, Vishweshwar Ghanakota, Alessandra Ferrari, Gregory N Nicola, Raffaele Bruno, Ignazio R Marino, Kenza Hamidouche, and Niraj K Jha. 2021. CovidDeep: SARS-CoV-2/COVID-19 test based on wearable medical sensors and efficient neural networks. *IEEE Trans. Consumer Electronics* 67, 4 (2021), 244–256.

[26] Shayan Hassantabar, Prerit Terway, and Niraj K. Jha. 2020. TUTOR: Training neural networks using decision rules as model priors. *arXiv preprint arXiv:2010.05429* (2020).

[27] Shayan Hassantabar, Zeyu Wang, and Niraj K. Jha. 2021. SCANN: Synthesis of compact and accurate neural networks. *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems* (2021).

[28] Katherine A. Herborn, James L. Graves, Paul Jerem, Neil P. Evans, Ruedi Nager, Dominic J. McCafferty, and Dorothy E.F. McKeegan. 2015. Skin temperature reveals the intensity of acute stress. *Physiology & Behavior* 152 (2015), 225–230.

[29] Deping Kuang, Xiaojiao Guo, Xiu An, Yilu Zhao, and Lianghua He. 2014. Discrimination of ADHD based on fMRI data with deep belief network. In *Proc. Int. Conf. Intelligent Computing*. 225–232.

[30] Deping Kuang and Lianghua He. 2014. Classification on ADHD with deep learning. In *Proc. Int. Conf. Cloud Computing and Big Data*. 27–32.

[31] Jill Fain Lehman. 2000. The diagnostic and statistical manual of mental disorders. *Am. Psychiatric Assoc.* (2000).

[32] Huijie Lin, Jia Jia, Quan Guo, Yuanyuan Xue, Qi Li, Jie Huang, Lianhong Cai, and Ling Feng. 2014. User-level psychological stress detection from social media using deep neural network. In *Proc. Int. Conf. Multimedia*. 507–516.

[33] David J. Malan, Thaddeus Fulford-Jones, Matt Welsh, and Steve Moulton. 2004. CodeBlue: An ad hoc sensor network infrastructure for emergency medical care. In *Proc. Int. Workshop Wearable and Implantable Body Sensor Networks*.

[34] Alban Maxhuni, Angélica Muñoz-Meléndez, Venet Osmani, Humberto Perez, Oscar Mayora, and Eduardo F. Morales. 2016. Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. *Pervasive and Mobile Computing* 31 (2016), 50–66.

[35] D. Mehrotra, D. Nagpal, R. Srivastava, and R. Nagpal. 2018. Analyse power consumption by mobile applications using fuzzy clustering approach. *Int. J. Engineering* 31, 12 (2018), 2037–2043.

[36] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T. Dudley. 2018. Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics* 19, 6 (2018), 1236–1246.

[37] David C. Mohr, Mi Zhang, and Stephen M. Schueller. 2017. Personal sensing: Understanding mental health using ubiquitous sensors and machine learning. *Review of Clinical Psychology* 13 (2017), 23–47.

[38] Jamie T. Mullins and Corey White. 2019. Temperature and mental health: Evidence from the spectrum of mental health outcomes. *J. Health Economics* 68 (2019), 102240.

[39] Patricia J. Murphy, Mark G. Frei, and Demitri Papolos. 2014. Alterations in skin temperature and sleep in the fear of harm phenotype of pediatric bipolar disorder. *J. Clinical Medicine* 3, 3 (2014), 959–971.

[40] J.T. O'Brien, P. Gallagher, D. Stow, N. Hammerla, T. Ploetz, M. Firbank, C. Ladha, K. Ladha, D. Jackson, Roisin McNaney, et al. 2017. A study of wrist-worn activity measurement as a potential real-world biomarker for late-life depression. *Psychological Medicine* 47, 1 (2017), 93–102.
[41] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. 2017. Predicting healthcare trajectories from medical records: A deep learning approach. *J. Biomedical Informatics* 69 (2017), 218–229.
[42] Walter H.L. Pinaya, Ary Gadelha, Orla M. Doyle, Cristiano Noto, André Zugman, Quirino Cordeiro, Andrea P. Jackowski, Rodrigo A. Bressan, and João R. Sato. 2016. Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia. *Scientific Reports* 6 (2016), 38897.
[43] Guilherme V. Polanczyk, Giovanni A. Salum, Luisa S. Sugaya, Arthur Caye, and Luis A. Rohde. 2015. Annual research review: A meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. *J. Child Psychology and Psychiatry* 56, 3 (2015), 345–365.
[44] Andrew G. Reece and Christopher M. Danforth. 2017. Instagram photos reveal predictive markers of depression. *EPJ Data Science* 6 (2017), 1–12.
[45] Farig Sadeque, Dongfang Xu, and Steven Bethard. 2017. UArizona at the CLEF eRisk 2017 pilot task: Linear and recurrent models for early depression detection. In *Proc. CEUR Workshop*, Vol. 1866. NIH Public Access.
[46] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. MobileNet v2: Inverted residuals and linear bottlenecks. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*. 4510–4520.
[47] Marco Sarchiapone, Carla Gramaglia, Miriam Iosue, Vladimir Carli, Laura Mandelli, Alessandro Serretti, Debora Marangon, and Patrizia Zeppegno. 2018. The association between electrodermal activity (EDA), depression and suicidal behaviour: A systematic review and narrative synthesis. *BMC Psychiatry* 18, 1 (2018), 1–27.
[48] Hugo G. Schnack, Mireille Nieuwenhuis, Neeltje E.M. van Haren, Lucija Abramovic, Thomas W. Scheewe, Rachel M. Brouwer, Hilleke E. Hulshoff Pol, and René S. Kahn. 2014. Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophrenia, bipolar disorder and healthy subjects. *Neuroimage* 84 (2014), 299–306.
[49] Kenneth O. Stanley and Risto Miikkulainen. 2002. Evolving neural networks through augmenting topologies. *Evolutionary Computation* 10, 2 (2002), 99–127.
[50] Zachary Steel, Claire Marnane, Changiz Iranpour, Tien Chey, John W. Jackson, Vikram Patel, and Derrick Silove. 2014. The global prevalence of common mental disorders: A systematic review and meta-analysis 1980–2013. *Int. J. Epidemiology* 43, 2 (2014), 476–493.
[51] Chang Su, Zhenxing Xu, Jyotishman Pathak, and Fei Wang. 2020. Deep learning in mental health outcome research: A scoping review. *Translational Psychiatry* 10, 1 (2020), 1–26.
[52] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. 2019. MnasNet: Platform-aware neural architecture search for mobile. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*. 2820–2828.
[53] Katarzyna Wac, Aart Van Halteren, and Dimitri Konstantas. 2006. QoS-predictions service: Infrastructural support for proactive QoS-and context-aware mobile services (position paper). In *Proc. Int. Conf. Move to Meaningful Internet Systems*. Springer, 1924–1933.
[54] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. 2019. FBNet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*. 10734–10742.
[55] Bichen Wu, Alvin Wan, Xiangyu Yue, Peter Jin, Sicheng Zhao, Noah Golmant, Amir Gholaminejad, Joseph Gonzalez, and Kurt Keutzer. 2018. Shift: A zero flop, zero parameter alternative to spatial convolutions. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*. 9127–9135.
[56] Yuankun Xue and Paul Bogdan. 2017. Constructing compact causal mathematical models for complex dynamics. In *Proc. Int. Conf. Cyber-Physical Systems*. 97–107.
[57] Yuankun Xue, Sergio Pequito, Joana R. Coelho, Paul Bogdan, and George J. Pappas. 2016. Minimum number of sensors to ensure observability of physiological systems: A case study. In *Proc. Allerton Conf. Communication, Control, and Computing*. 1181–1188.
[58] Yuankun Xue, Saul Rodriguez, and Paul Bogdan. 2016. A spatio-temporal fractal model for a CPS approach to brain-machine-body interfaces. In *Proc. Design, Automation & Test in Europe Conf. & Exhibition*. 642–647.
[59] Zikun Yang, Paul Bogdan, and Shahin Nazarian. 2021. An in silico deep learning approach to multi-epitope vaccine design: A SARS-CoV-2 case study. *Scientific Reports* 11, 1 (2021), 1–21.
[60] Hongxu Yin, Ayten Ozge Akmandor, Arsalan Mosenia, and Niraj K. Jha. 2018. Smart healthcare. *Foundations and Trends in Electronic Design Automation* 12, 4 (2018), 401–466.
[61] Hongxu Yin, Bilal Mukadam, Xiaoliang Dai, and Niraj K. Jha. 2019. DiabDeep: Pervasive diabetes diagnosis based on wearable medical sensors and efficient neural networks. *IEEE Trans. Emerging Topics in Computing* (2019).
[62] Ling-Li Zeng, Huaning Wang, Panpan Hu, Bo Yang, Weidan Pu, Hui Shen, Xingui Chen, Zhening Liu, Hong Yin, Qingrong Tan, et al. 2018. Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity MRI. *EBioMedicine* 30 (2018), 74–85.
[63] Barret Zoph and Quoc V. Le. 2017. Neural architecture search with reinforcement learning. In *Proc. Int. Conf. Learning Representations*.