Deep Physiological Arousal Detection in a Driving Simulator using Wearable Devices

Aaqib Saeed*
University of Twente
Enschede, Netherlands
a.saeed@student.utwente.nl

Stojan Trajanovski
Philips Research
Eindhoven, Netherlands
stojan.trajanovski@philips.com

Maurice van Keulen
University of Twente
Enschede, Netherlands
m.vankeulen@utwente.nl

Jan van Erp
University of Twente
Enschede, Netherlands
jan.vanerp@utwente.nl

***Abstract*—Driving is an activity that requires considerable alertness. Insufficient attention, imperfect perception, inadequate information processing, and sub-optimal arousal are possible causes of poor human performance. Understanding of these causes and the implementation of effective remedies is of key importance to increase traffic safety and improve driver's well-being. For this purpose, we used deep learning algorithms to detect arousal level, namely, under-aroused, normal and over-aroused for professional truck drivers in a simulated environment. The physiological signals are collected from 11 participants by wrist wearable devices. We presented a cost effective ground-truth generation scheme for arousal based on a subjective measure of sleepiness and score of stress stimuli. On this dataset, we evaluated a range of deep neural network models for representation learning as an alternative to handcrafted feature extraction. Our results show that a 7-layers convolutional neural network trained on raw physiological signals (such as heart rate, skin conductance and skin temperature) outperforms a baseline neural network and denoising autoencoder models with weighted F-score of *0.82* vs. *0.75* and Kappa of *0.64* vs. *0.53*, respectively. The proposed convolutional model not only improves the overall results but also enhances the detection rate for every driver in the dataset as determined by leave-one-subject-out cross-validation.**

***Index Terms*—Arousal Detection, Deep Learning, Driving Simulator, Sleepiness, Fatigue, Stress, Convolutional Neural Network, Wearable Devices.**

I. Introduction

Driving is a complex task involving several motor and cognitive abilities. Inadequate human performance is a major cause of road traffic accidents. Imperfect perception, insufficient attention, inadequate information processing, and sub-optimal arousal is mentioned as possible causes for poor human performance. For instance, driver drowsiness or fatigue caused by extended hours of driving, as well as situations of cognitive overload, can significantly impair a driver's ability to react appropriately to relevant events [1], [2]. Understanding of these causes and the implementation of effective remedies is of key importance to increase traffic safety and driver well-being..

The physiological arousal level can be described as the available capacity to perform the task in timely and effective manner. The potential threat of both under-arousal as well as over-arousal is reflected in many human performance models (e.g. see [3] for an overview). The more complex models take at least the relationship between task demands, workload, effort and performance into account. Among them, workload is considered a multidimensional, multifaceted concept that is difficult to define and quantify using a single representative measure [4]. However, in the context of driving a simple model consisting of a single dimension —here referred to as arousal— may suffice.

Over a century ago, Yerkes and Dodson [5] established a law stating that the relationship between performance and level of arousal has an inverted U-shape (see Fig. 1). If physiological signals reliably reflect a possible threat of under-arousal or over-arousal before a decline in driving performance becomes noticeable, they may form the basis for an effective remedy. The rapid development of wearable sensors to record physiological parameters; over the past years makes the development of effective solutions more feasible than ever before. Moreover, the solutions built by leveraging raw signals collected from non-invasive wearable devices, such as Photoplethysmogram (PPG) sensor for heart rate, is more feasible to use in an everyday situation than an Electroencephalography (EEG) and Electrocardiography (ECG). Another reason for using physiological signals as opposed to driving behaviour and vehicle data is that they are found to be more indicative of driver's state as compared to driving behaviour [6].

In earlier research, significant work has been done to detect stress and fatigue using machine learning and signal processing methods, on both simulator and on-road datasets [7]. These

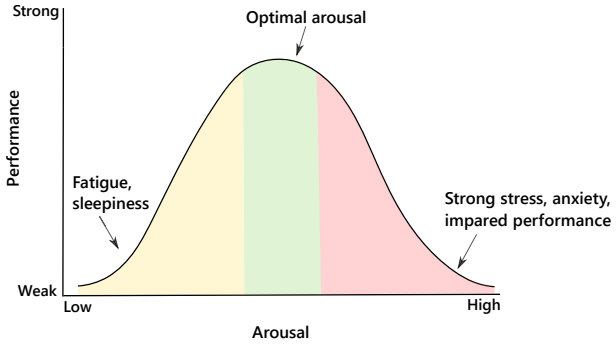


Fig. 1. Illustration of Yerkes-Dodson law [1].

* Corresponding author. This work is done while A. S. was an intern at Philips Research, Eindhoven.

proposed techniques for driver state detection mainly rely on hand-crafted features to classify physiological signal segments (e.g. as either stressed or fatigued). The process of manual feature engineering is cumbersome and requires extensive domain knowledge. Furthermore, the generated features are not guaranteed to be optimally discriminant to solve the task at hand and hence require usage of feature selection or dimensionality reduction techniques. Several recent studies have shown that better performance can be achieved when feature extraction is performed jointly along with training models in an end-to-end fashion. For instance, Sutskever, Vinyals and V. Le [8] proposed an approach for sequence learning to extract discriminant features for machine translation task. Therefore, end-to-end learning via deep learning algorithms has the potential to have a significant impact on problems involving multivariate time series datasets. It can substitute manually designed feature extraction procedures and deep models can automatically learn variations and trends in the signal.

The main contributions of this paper consist of the following:

- Tackling stress and drowsiness together as a problem of physiological arousal detection, using only raw data collected from wearable devices.
- A ground truth generation scheme for physiological arousal by combining self-assessment questionnaire of sleepiness and score of task-induced stimuli from a stressful task.
- Finding a robust deep neural network architecture for arousal classification.
- Empirically exploring the effect of techniques to resolve data imbalance with a deep neural network.

The rest of the paper is structured as follows: Section II presents the background and related studies for stress and fatigue detection in professional drivers. The experimental setting for data collection, arousal ground truth generation, and signal segmentation is provided in Section III. The arousal classification problem formulation, explanation of widely used deep neural networks, and model architectures are presented in Section IV. Subsequently, the results of performed experiments are provided in Section V. The paper is concluded in Section VI by highlighting the main findings, discussing limitations of the current work and providing directions for future research.

II. Related Work

Many physiological measures correlate with specific mental or cognitive state [9]; amongst others, based on activity patterns of brain, heart, skin conductivity and eyes [10]. However, as comprehensively discussed by Fairclough [11], the relationship between physiological measures and psychological meaning is complex. Here, we will focus on the general patterns related to underload and overload (or under-arousal and over-arousal) and on brain patterns, while later on we will discuss studies concerned with driving in more detail with a focus on heart and skin conductance indices.

Indices for underload are mainly based on the theta and alpha power in the EEG. Generally, increased theta power correlates with poor performance in sleep deprived subjects [12] and in vigilance tasks [13]. Lal and Craig [14] concluded, based on an extensive review, that changes in theta and delta activity are strongly linked to transition to fatigue. Stampi, Stone and Michimori [15] showed the usefulness of alpha activity as an index for sleepiness. More precisely, they validated the Alpha Attenuation Test (AAT) as an index for sleepiness with sleep deprived subjects. The AAT is based on the observation that, when operators get sleepier, alpha activity with eyes open increases and decreases with eyes closed. Other valid indices for underload may be based on heart rate and Heart Rate Variability (HRV). Generally, a drop in heart rate or an increase in HRV can occur at the beginning of a drowsiness state [2], [14]. It is important to notice that both indices are influenced by a variety of factors, including physical movement, mental activity, and emotional state [16]. Finally, and far less explored, body temperature is also linked to arousal as it reflects on the person's state, reflecting the autonomic responses [17].

Driving studies are also relevant, because cognitive studies do not necessarily generalize and may be different from the stationary states (see [18] for a review).

1) Under-aroused (Underload): Several authors used simulation environments to investigate the physiological effects of sleep deprivation or long hours of continuous driving. Lal and Craig [19] showed that delta and theta activity increased during fatigue, and heart rate decreased. Zhao, Chunlin, et al. [20] reported significant changes of EEG alpha and beta power, of the P300 amplitude, of the approximated entropy of the ECG, and of the lower and upper bands of HRV power before and after 90 minutes driving. Liang, Wen Chieh, et al. [21] found decreased heart rate, systolic pressure, LF/HF, and palm temperature; and increased HRV and parasympathetic indices HF(AU) and HF(NU) after 120 minutes of simulated driving. Other authors analysed physiological recordings taken during real driving. Raggatt and Morrissey [22] report a lower heart rate after 9 to 12 hours of driving as did [23] for twelve train drivers during monotonous stretches. Opposite effects on heart rate were reported by Apparies, Riniolo and Porges [24], who followed truck drivers on a route that lasted between 8 and 10 hours and found increased heart rate and decreased HRV. Under conditions of increasing sleep deprivation (up to 34 hours), Furman, G. Dorfman, et al. [25] report on 59 falling asleep (FA) events. The mean heart rate and overall variability decrease during FA events by 2.2 SD and 2.9 SD below regional means.

2) Over-aroused (Overload): The patterns reported for (over) load are generally more consistent across studies than those reported above for under load. Mehler, Bruce, et al. [6] examined the sensitivity of heart rate, skin conductance, and respiration rate as measures of the (over) load in a simulated driving environment. Heart rate, skin conductance, and res-

piration rate increased with increasing task demand. Similar results were found for heart rate and skin conductance [26] and these patterns are similar for simulated and real driving [27], [28]. What we should notice, though, is that the majority of the studies on (over) load employ a secondary cognitive task to vary task load. Brouwer, Dijksterhuis, and van Erp. [29] points out that the physiological effects of mental effort as manipulated through cognitive task difficulty differ from effects of mental effort as manipulated through a visuomotor task such as lane keeping in a simulated driving. Most notably [30], [31] demonstrate that, the heart rate, heart rate variability, and respiration may not be affected by task difficulty of visuomotor tasks like driving.

III. Dataset and Methodology

A. Data Collection Protocol

We collected heart rate, skin conductance, skin temperature and accelerometer data from 11 participants (professional truck drivers) using wrist-worn devices. The heart rate with a frequency of one Hz was derived from PPG sensor data and other physiological signals were recorded at a frequency of 10 Hz. The experiment was realized with driving simulation software and participants received standardized instructions from an audiotape. Two experimental factors were manipulated: namely stress and sleepiness. The high stress was induced by means of secondary arithmetic subtraction task. It is a component of widely used Trier Social Stress Test [32], where a user has to perform serial subtraction verbally in a loud manner and have to start over from the last correct answer, if a mistake is made. Likewise, in a fatigue trial, Karolinska Sleepiness Scale (KSS) was used for evaluating subjective sleepiness of each driver. It spans nine levels and asks the user to provide the number that most closely represents their sleepiness level at the moment. The KSS appears to be most widely used a sensitive and reliable measure of sleepiness [33]. Moreover, studies show a significant correlation between the KSS and objective measures of driving performance such as standard deviation of the lateral position and blink duration [34]. This makes it a feasible measure to derive ground truth for supervised models as compared to video-coding, which requires substantial human effort and have high chances of bias being included in the generated labels.

The experiment consisted of six major steps that are given in Fig. 2. Before the start of the experiment, participants filled-in the KSS and other related forms. The first experimental trial consisted of normal driving (baseline condition) for 15 minutes. Afterwards, each subject was asked twice to count $1-60$ as a moderate stress activity with a very small interval between the two activities. After a one-minute period of normal driving and to induce high stress, the subject was asked to count backward from a random number in steps of 7 in approximately 30 seconds. After that the subject was asked to count backward again from another random number. The process was repeated for approximately 5 minutes. The length of the stress simulation task was 25 minutes, including baseline. In the break period (which varied from driver to driver), participants filled-in the KSS form for the second time. Then, the second experiment (named sleepiness or fatigue) phase started and lasted 90 minutes. In this trial, no secondary tasks were applied. Every 10 minutes, a KSS prompt was given (on tablet) to the drivers to collect their sleepiness level. At the end of the experiment, drivers filled-in the KSS and other required forms, and devices were removed.

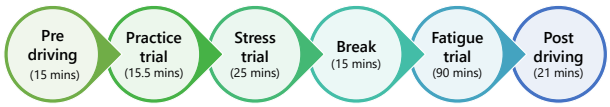


Fig. 2. Sequence and duration of events in a simulator study.

B. Ground Truth Annotation

To derive the ground truth labels for arousal, we followed the experimental protocol and used stress and KSS ratings. The data collected during baseline, moderate and high-stress trial was assigned labels of 1, 2 and 3, respectively. Moreover, the data points during instruction periods were simply assigned the label of zero to avoid wrong labeling. During a 90 minutes fatigue trial, drivers were asked every 10 minutes for a KSS score. Furthermore, the two KSS scores provided by the drivers at the start of the experiment and before the break were also used. The values are linearly interpolated between the start and the break, to get a discrete range of KSS scores that we used for arousal ground truth labeling. The KSS scores from 1 to 5 were considered to be in "alert" state, whereas, 6 to 9 were considered as "sleepy" state.

Let s denote the stress label and k represents the set of KSS scores. Then the arousal label l can be determined as follows:

$$l = \begin{cases} \text{under-aroused, if } s \in \{1,2,3\} \text{ and } k \in \{6,7,8,9\} \\ \text{normal, if } s \in \{1,\ 2\} \text{ and } k \in \{1,2,3,4,5\} \\ \text{over-aroused, if } s = 3 \text{ and } k \in \{1,2,3,4,5\} \end{cases}$$

C. Pre-processing and Segmentation

Physiological signals vary significantly from person to person and depend on several factors such as age and diet etc. [35]. Hence, it becomes important to minimize this variation. We minimally preprocess the dataset to let deep nets extract key non-linear features. The biometric signals of each driver are mean normalized by baseline to have zero mean and unit-variance [see equation (1)]. The mean ($\overline{x_b}$) and standard deviation (σ_b) is calculated from the normal baseline driving of 15 minutes. The day-to-day variations that can be caused by several different factors, such as, mood fluctuations are not considered in this work as the total duration of the simulation task was approx. two hours.

$$x' = \frac{x - \overline{x_b}}{\sigma_b} \tag{1}$$

It is mentioned earlier that, heart rate and other physiological signals had a different sampling rate. We upsampled the heart rate using linear interpolation to match the frequency of the rest to 10 Hz. The upsampling is performed to keep

the dataset size large enough and avoid losing meaningful information. Likewise, we used sliding window approach to extract signal segments with fixed step size of 10 seconds. The windows of 10, 30, 60 and 90 seconds (each having 100, 300, 600 and 900 samples respectively) are considered to find the optimal window size. However, in the literature [36], [37], a window of 60 and 300 seconds is usually used for skin conductance and HRV feature extraction, respectively. Moreover, to assign a corresponding class label y_i to a segment i; its value is taken to be the mode of class labels L within the window, which corresponds to a arousal level category.

IV. AROUSAL CLASSIFICATION

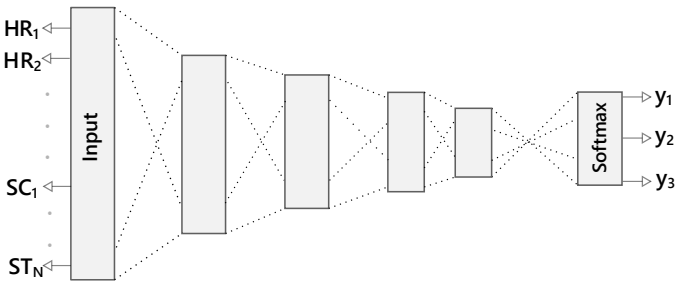
In the following section, we first briefly explain the problem definition followed by a short explanation of deep neural networks. We then describe architectural choices including how to represent the raw physiological signal input for deep nets. Finally, several specific design possibilities and details of techniques for handling data imbalance are specified.

A. Problem Definition

We considered the problem of arousal detection as a supervised sequence classification. In this task, the objective is to assign a single label to an input sequence. This problem can formulated as follows, let $\left\{(\boldsymbol{x_i}, y_i)\right\}_{i=1}^{N}$ be a dataset with N training examples. Each example pair (x_i, y_i) can be thought of as a pair of physiological signal vector $\boldsymbol{x_i}$ along with corresponding label y_i assigned to the sequence, where $\boldsymbol{x}_i$ is a sequence $\boldsymbol{x_i} = (x^1, x^2, x^3, \ldots, x^t)^T, x \in R^d$. For a new test sequence $\hat{\boldsymbol{x}}$, the goal is to predict a label for an entire sequence. Generally, such problems are solved by a) performing high-level feature extraction from sequences and using traditional classifiers such as SVM. b) employing distance based algorithms such as Dynamic Time Warping or c) probabilistic methods like Hidden Markov Model (HMM). In this work, we employed deep learning methods that incorporate characteristics of antecede techniques. For example, convolutional neural network hierarchically learns complex nonlinear features composed of an earlier ones and act as a replacement for handcrafted feature engineering. The rest of the section explains deep nets working on physiological signals in detail.

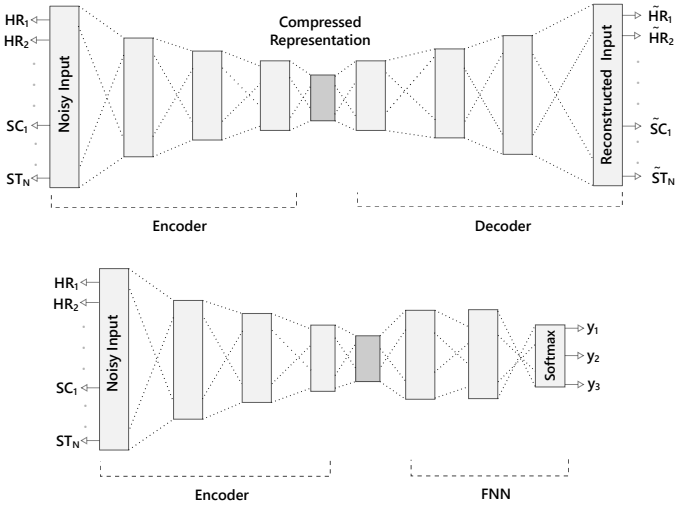
B. Neural Network and Denoising Autoencoder

We used plain 4-layers Neural Network (NN) and Denoising Autoencoder (DAE) as our baseline models. The fully connected layers of the NN model had 256, 128, 64 and 32 neurons, respectively. The $sigmoid$ function is applied as a nonlinearity, whereas, the $softmax$ function is used in the last layer to get normalized output probabilities. In the second baseline model, the DAE is used for unsupervised pre-training to learn initial signal representation. The architecture of the DAE was similar to that of the NN, where, $dropout$ [38] is used to introduce noise in the input with a probability of 0.2, and $L2$ penalty is used on the weights of the encoders. Instead



* HR : Heart Rate, SC: Skin Conductance, ST: Skin Temperature

Fig. 3. Baseline neural network architecture.



* HR : Heart Rate, SC: Skin Conductance, ST: Skin Temperature, FNN: Feed-Forward Neurar Network

Fig. 4. Baseline architecture of denoising autoencoder for unsupervised pre-training with fully connected neural network.

of $sigmoid$, $softsign$ is used as nonlinearity in the encoder and the decoder of the network. Afterward, the decoder is replaced with 2-layer FNN with $sigmoid$ as nonlinearity, resulting in a six layers feed-forward network, which is trained end-to-end. The input data fed into the network corresponds to a flattened vector of physiological signals. Each input segment $\boldsymbol{x_i} \in \mathbb{R}^{t*s}$ was of the selected windows size t (e.g. 60 seconds) extracted using sliding window for biometric signal s concatenated together. The NN model architecture is depicted in Fig. 3 and the DAE model with fully connected layers is shown in Fig. 4. The mean squared error and negative log-likelihood are used as objective functions for unsupervised and supervised model training, respectively.

C. Convolutional Neural Network

Convolutional neural networks (CNN) are generally used as feature extractors on various types of data including images, text, and time-series. They have the capability to learn local salient features by weight sharing while applying mathematical operation called "convolution" along with non-linearities. Mostly, in CNN an operation called "pooling" is also employed after convolution block. It performs sub-sampling by

replacing particular values within a subregion by a single value. Thus, introducing translational invariance to the model. For further details see [39], [40].

We represent the input for CNN model as a multidimensional array with the number of time steps as the width and the number of physiological signals as input features. Let n be the total number of examples (after signal segmentation), t be the number of samples or time steps and s denote the number of physiological parameters, the resulting input tensor X will be of shape $n \times t \times s$. We first applied depthwise convolution to extract individualistic features from each physiological signal. It is performed independently over each input channel i.e. heart rate, skin conductance etc. Subsequently, several temporal convolutions and $average$ pooling operations are applied to learn a wide range of complex features. The CNN model had three convolution-average-pooling blocks, with a depthwise convolution in the first stage followed by standard convolution-average-pooling, densely connected layer having 512 neurons and a $softmax$ classification layer (see Fig. 5). The exponential linear unit [41] ($elu = exp(x) - 1$ if $x \leq 0$; x otherwise) is used as nonlinear activation function in convolution layers, whereas, $sigmoid$ activation is used in a densely connected layer. The $L2$ regularization is used on the weights of the convolution layers and dropout with a probability of 0.2 is used on a densely connected layer to avoid over-fitting and improve model generalization.

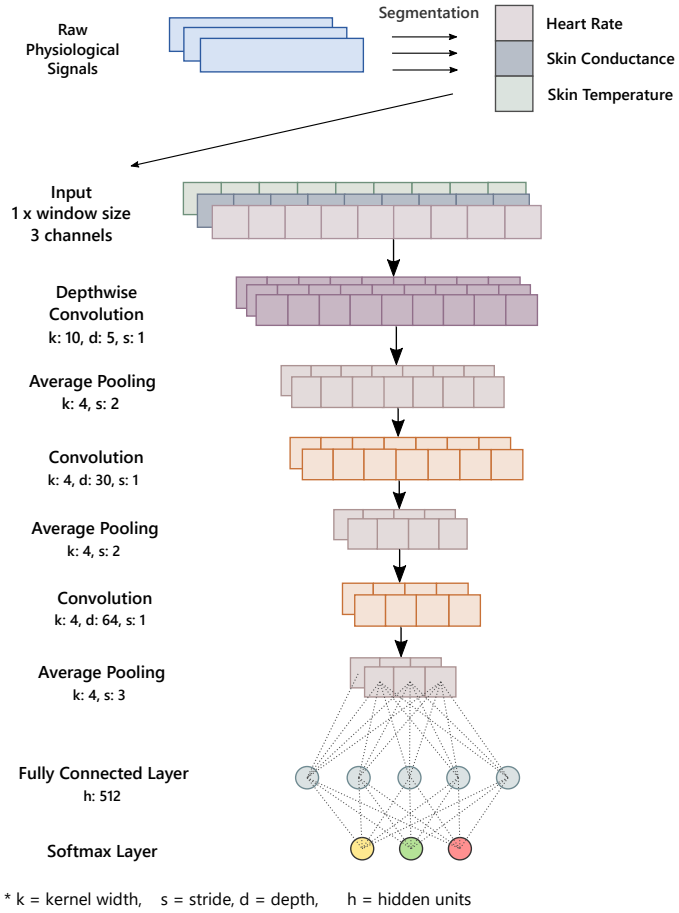


Fig. 5. Convolutional neural network architecture.

D. Recurrent Neural Network

Recurrent neural network is a connectionist model that has the capability to capture sequential time dependencies in the input data. It can preserve the state from an arbitrarily large window and overcome major limitations of the standard neural network. In this work, we employed Gated Recurrent Unit (GRU) [42] based RNN model. The hidden state dimension was 128 and retained across batch training iterations (or also called stateful). In GRU based model, the hidden state h of the RNN can be formulated as given by equation (2). The input data fed into RNN was of the same dimensions as CNN model i.e. $X \in \mathbb{R}^n \times \mathbb{R}^t \times \mathbb{R}^s$. The last output from the model is fed into a fully connected layer with 256 neurons and $tanh$ activation function, which later passed as input to $softmax$ layer.

$$
\begin{aligned}
i_f &= \sigma(W^{ff} \times x_t + W^{fr} \times h^{t-1} + b_f) \\
i_o &= \sigma(W^{of} \times x_t + W^{or} \times h^{t-1} + b_o) \\
\widetilde{h}^t &= \tanh\left(W^{yf} \times x^t + W^{yr} \times (i_f \odot h^{t-1}) + b_{\widetilde{h}^t}\right) \\
h^t &= i_o \odot h^{t-1} + (1 - i_o) \odot \widetilde{h}^t
\end{aligned}
\tag{2}
$$

E. Hybrid Models

The convolutional and recurrent models capture a local and global view of the data, respectively. A natural question to ask, whether these two models can be combined into a unified model for learning both local representation and long-range dependencies in the signals. Therefore, we also created two hybrid architectures by integrating CNN and RNN together (see Fig. 6). We took inspiration for second hybrid architecture from work on learning representation from EEG [43]. In the first variant, we dropped $softmax$ and fully connected layer of CNN model and fed last layer output into the RNN. This can be seen as having a recurrent connection over an entire window, instead of on each sample. The output from the last stage of the recurrent layer is fed into a fully connected layer having 512 neurons, which then pass on to $softmax$ for classification. In the second variation, RNN input was the same as first version but we added another convolution layer, to convolve over existing learned features from CNN. Afterward, the last output from recurrent layer along with the output of new convolution layer was concatenated and fed into a fully connected layer. These hybrid models were trained from scratch without using any trained weights from existing models.

F. Tackling Data Imbalance

Generally, supervised learning techniques work well with reasonably balanced datasets, where, a representation of each class is uniform [44]. In our case, the over-aroused class

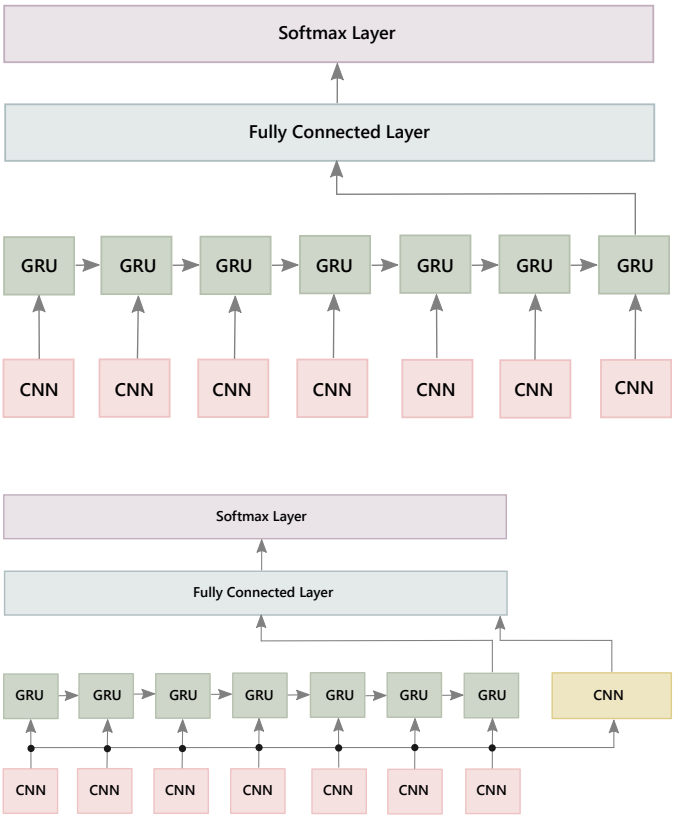


Fig. 6. Attempted hybrid architectures of convolutional and recurrent neural networks.

is underrepresented due to a short duration of the stressful task as compared to sleepiness trial. To resolve this, we applied over-sampling, threshold-moving and cost sensitive loss function techniques and performed the experiments with convolutional neural network model discussed in Section IV-C. The mentioned techniques for resolving class imbalance are briefly reviewed here.

1) Threshold-Moving: The idea behind threshold-moving is to first define a cost matrix, which describes the misclassification cost of assigning an instance from one class to another. The normalized probabilities from the deep net are then modified according to equation (3) and class label with maximum probability is selected [44]. By using this method, network architecture and training procedure is not modified in any way, and the cost sensitivity is introduced during evaluation phase which can be seen as an advantage.

$$\hat{y_j}^* = \frac{\sum_{i=1}^{NC} y_j \times CM[j,i]}{C} \tag{3}$$

Where y_j are original output probabilities from a $softmax$ layer, NC is total number of classes, CM is a cost matrix, C is a sum of costs for each class and $\hat{y_j}^*$ and act as normalization term to keep $0 \leq \hat{y_j}^* \leq 1$.

2) Over-Sampling: Synthetic Minority Over-sampling Technique (SMOTE) is an oversampling technique that randomly generates additional data points by interpolation from the minority class samples. This method directly changes the distribution of the dataset to have an equal number of examples for every class and proved to be effective in learning from imbalanced data [45]. Moreover, in order to minimize the class overlap, a data cleansing approach i.e. *Tomek links* [46] is generally applied. It removes congested borderline data points of opposite classes that have minimal distance between them.

3) Weighted Categorical Cross Entropy: The deep neural network can be enabled to learn imbalance property of the data by having a cost sensitive loss function. The categorical cross entropy loss function is modified as proposed by [47] to reflect equal error from both majority and minority classes. The prior class probabilities are incorporated into a categorical cross entropy [see equation (5)], resulting in a modified objective function given by equation (4):

$$-\frac{1}{NC \times n} \sum_{i}^{n} \sum_{k}^{m} y_{i,k} \frac{\log \hat{y}(x_i, y_i; W)}{p(y_i)} \tag{4}$$

Where NC shows the number of classes, n is the batch size, m represents a total number of classes and $p(y_i)$ denotes the prior probability of the class. For detailed mathematical treatment of the loss function please see [47].

V. Results

The model evaluation strategy and empirical findings of our experiments are presented in this section.

The processed segmented data of each driver was randomly split in a stratified manner. 80% of the data were used for model training and validation, whereas, the remaining 20% were hold-out for final model evaluation. To determine the performance of the learned model for each driver, we used Leave-One-subject-Out Cross-Validation (LOOCV). In this method, the model is trained from scratch holding out data of one driver as validation set and using the rest as a training set. Weighted F-score and Kappa measures are calculated for LOOCV and hold-out test set as they are effective performance measures; when data is imbalanced.

In order to apply SMOTE on the current multiclass problem, we considered samples from majority classes as belonging to one class. For example, for generating samples of the over-aroused class, we combined data points of under-aroused and normal in one class. Likewise, in order to avoid generating a large number of synthetic data points that are not really representative of actual data points, we oversampled the over-aroused and normal classes by 25% and 50%, respectively. It is essential to note that the data is oversampled before the signal segmentation process and the smoted data is used only for training while the non-smoted is used for evaluation in LOOCV. Likewise, for threshold-moving, we experimented with four different cost matrices. Table I shows the cost matrix that provided better results with the CNN model as compared to the others.

We adopted Xaiver method proposed by [48] for initialization of models' weights. Moreover, negative log-

TABLE I
OPTIMAL COST MATRIX.

		j		
		Under-aroused	**Normal**	**Over-aroused**
i	**Under-aroused**	0	1	8
	Normal	1	0	9
	Over-aroused	1	1	0

likelihood[1] [39] is used as an objective function for multiclass classification problem, it is given by equation (5).

$$L = -\frac{1}{n}\sum_{i}^{n}\sum_{k}^{m} y_{i,k}\ log(\hat{y}_{i,k}) \tag{5}$$

Where n represent number of training examples, m denotes the number of classes, y_i is the true label and $\hat{y_i}$ is the model's output.

We first evaluated baseline models as their classification performance is compared with other, more complex architectures. The cross-validation results of various deep models trained on physiological signal segments of different window sizes are summarised in **Table II**. The NN model reached average validation F-score and Kappa of 0.75 and 0.53 respectively, for a window size of 30 seconds. Likewise, the pre-training using denoising autoencoder with two additional layers used for supervised training achieved F-score 0.76 and 0.55 Kappa for 30 seconds window size. These results show some improvement over a plain neural network possibly due to more layers; in addition to pre-training of the model.

The proposed CNN architecture outperformed baseline models, recurrent GRU and some hybrid architectures; and significantly improved results across drivers. This suggests that 1-D convolutions are able to extract important features in physiological signals than shallow baseline models which do not take temporality into account. It achieves values of 0.82 for the F-score and 0.64 for Kappa during cross-validation. Likewise, on a hold-out test set to attain 0.81 F-score and 0.64 Kappa for 60 seconds window size. To find the optimal architecture, we explore several specific design choices such as activation functions and pooling type (such as max pooling, ReLU activation etc.). We selected the architecture configuration that gave us an overall improvement on evaluation metrics (with low standard deviation) but also improved results for each driver. The reason that CNN is superior over RNN is in the the very large input sequence length and the recurrent connection on a sample by sample basis which slowed the learning in recurrent model, considerably.

The hybrids of convolution and recurrent networks are evaluated to overcome the issue of having a recurrent connection over each sample. In the first variant, the learned features from the convolution model are fed into the RNN followed by fully connected layer. In the second variation, in addition to feeding CNN features into RNN, we also fed those features into a fully connected layer. The hybrid architectures did not perform significantly better than the proposed CNN model (F-score 0.82 of CNN vs 0.744 and 0.747 of hybrid a and b respectively) and improved very little over baseline models.

The techniques to solve data imbalance are evaluated using the optimal CNN model. The goal was to see if the detection rate of minority classes can be improved. The overall averaged results are not improved but the detection rate for the over-aroused class was slightly higher when over-sampled (SMOTE) dataset was used with the CNN model. However, we believe that a possible reason for the limited improvement in a case of SMOTE, is that synthetic data points also have a class overlap issue and they do not truly represent training examples. Likewise, threshold-moving also did not improve much, as finding an optimal cost matrix can be seen as an optimization process in itself. Furthermore, the weighted categorical cross entropy loss function is assessed to handle data imbalance during a model learning process. We noticed that a batch size greater than 15 is feasible for optimization to begin smoothly due to the division term involved in the formula. The results achieved using this loss function did not improve much on the earlier achieved results using negative log-likelihood.

VI. DISCUSSION & CONCLUSION

The goal of this paper was to apply widely used deep learning algorithms as an alternative to handcrafted feature extraction for arousal classification. The manual feature engineering is a requirement of traditional machine learning algorithms, which is a cumbersome process and limited by the researcher's ability to create discriminant features. Moreover, we used physiological parameters of professional truck drivers collected from wearable devices during a simulation study. We presented a ground truth generation scheme for arousal based on a subjective measure of sleepiness and a score of stress stimuli. This scheme is cost effective and efficient as compared to using video decoding for ground truth labels generation.

The convolutional neural network trained on raw physiological signals (i.e. heart rate, skin conductance and skin temperature) outperformed baseline neural network and denoising autoencoder models with a weighted F-score of 0.82 vs. 0.75 and a Kappa of 0.64 vs. 0.53, respectively. Moreover, the Convolutional Neural Network (CNN) outperform Gated Recurrent Unit (GRU) and hybrid models of CNN + GRU/dense layers. The proposed convolutional model not only improve the overall results, but enhanced the detection rate for every driver in the dataset as determined by leave-one-subject-out cross-validation. Likewise, several specific design choices were evaluated for a convolutional neural network to find a robust architecture, we found elu activation and $average$ pooling to give optimal results as compared to other configurations. Due to the short duration of the stressful task and a varying degree of drowsiness, the dataset was imbalanced. We empirically explored three methods for resolving this imbalance in the dataset. Using SMOTE, the synthetic

[1]Also known as Categorical Cross Entropy.

TABLE II
SUMMARISED RESULTS OF THE ATTEMPTED DEEP NEURAL NETWORK ARCHITECTURES FOR DIFFERENT WINDOW SIZES.

Model	Window Size	Validation F-score	Validation Kappa	Test F-score	Test Kappa
Neural Network	10	0.749	0.517	0.757	0.526
	30	**0.757**	**0.531**	**0.757**	**0.535**
	60	0.740	0.517	0.739	0.522
	90	0.747	0.538	0.760	0.568
Denoising Autoencoder	10	0.736	0.524	0.748	0.538
	30	**0.762**	**0.558**	**0.763**	**0.563**
	60	0.763	0.549	0.761	0.553
	90	0.729	0.548	0.727	0.529
Convolutional Neural Network (CNN)	10	0.766	0.551	0.761	0.547
	30	0.801	0.627	0.799	0.628
	60	**0.821**	**0.642**	**0.817**	**0.649**
	90	0.827	0.640	0.814	0.625
GRU - Recurrent Neural Network	10	0.685	0.416	0.678	0.422
	30	0.689	0.448	0.684	0.445
	60	**0.731**	**0.449**	**0.714**	**0.476**
	90	0.689	0.462	0.679	0.454
Hybrid Architecture - A	10	0.704	0.464	0.688	0.431
	30	0.735	0.547	0.725	0.545
	60	0.718	0.564	0.705	0.557
	90	**0.744**	**0.610**	**0.735**	**0.590**
Hybrid Architecture - B	10	0.742	0.490	0.706	0.453
	30	0.738	0.544	0.723	0.526
	60	**0.740**	**0.570**	**0.740**	**0.570**
	90	0.747	0.595	0.738	0.559
CNN - SMOTE	10	0.749	0.554	0.733	0.524
	30	0.775	0.610	0.770	0.610
	60	**0.770**	**0.630**	**0.770**	**0.630**
	90	0.787	0.649	0.773	0.621
CNN - Threshold Moving	10	0.729	0.526	0.740	0.537
	30	0.757	0.552	0.767	0.570
	60	0.796	0.596	0.802	0.616
	90	**0.810**	**0.617**	**0.804**	**0.606**
CNN - Weighted Categorical Cross Entropy	10	0.690	0.480	0.693	0.484
	30	0.700	0.526	0.686	0.507
	60	**0.736**	**0.585**	**0.749**	**0.601**
	90	0.733	0.588	0.721	0.567

data generation improved the detection rate of a convolutional neural network on over-aroused class but only slightly.

The major limitation of the presented models is the difficulty in differentiating between normal and over-arousal data points. One rationale could be because of the high class overlap between normal and over-aroused classes. More specifically, normal arousal data points were mostly comprised of baseline and no drowsiness signal during the sleepiness trial; hence they shared a similar characteristic with the over-aroused class. Likewise, as the physiological signals show huge interpersonal variation. For this reason, we suggest that the initial model should be personalized for each driver by training on one's newly collected data after deployment in production.

The purpose of the developed model should be nudging the drivers to improve their alertness and safety. Therefore, the model can be deployed locally on a smartphone for use in a real-life situation; while preserving privacy. It provides an opportunity to update an initial global model by aggregating local models of various drivers. The federated learning [49] approach can be naturally applied to this case as it takes non-IID and unbalanced nature of training data into account. However, getting the ground truth labels for arousal can be tricky in such a situation. We suggest using subjective measures, for instance, asking the user explicitly or looking in-combination with behavioural data like app usage to determine the correct labels. Moreover, another important future direction could be to understand the decision-making process of deep models trained on time-series datasets. Lastly, generative algorithms such as Generative Adversarial Networks [50] and Variational Autoencoder [51] are worth exploring, especially for data generation to solve the class imbalance problem.

REFERENCES

[1] Joseph F Coughlin, Bryan Reimer, and Bruce Mehler. Driver wellness, safety & the development of an awarecar. *AgeLab, Mass Inst. Technol., Cambridge, MA*, 2009.

[2] José Vicente, Pablo Laguna, Ariadna Bartra, and Raquel Bailón. Detection of driver's drowsiness by means of hrv analysis. In *2011 Computing in Cardiology*, pages 89–92. IEEE, 2011.

[3] Jan BF van Erp, Hans JA Veltman, and Marc Grootjen. Brain-based indices for user system symbiosis. In *Brain-Computer Interfaces*, pages 201–219. Springer, 2010.

[4] Daniel Gopher and Emanuel Donchin. Workload: An examination of the concept. 1986.

[5] Robert M Yerkes and John D Dodson. The relation of strength of stimulus to rapidity of habit-formation. *Journal of comparative neurology and psychology*, 18(5):459–482, 1908.

[6] Bruce Mehler, Bryan Reimer, Joseph Coughlin, and Jeffery Dusek. Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *Transportation Research Record: Journal of the Transportation Research Board*, (2138):6–12, 2009.

[7] Nandita Sharma and Tom Gedeon. Objective measures, sensors and computational techniques for stress recognition and classification: A survey. *Computer methods and programs in biomedicine*, 108(3):1287–1301, 2012.

[8] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[9] Jan BF van Erp, Anne-Marie Brouwer, and Thorsten O Zander. Using neurophysiological signals that reflect cognitive or affective state. *Frontiers in neuroscience*, 9, 2015.

[10] Maarten A Hogervorst, Anne-Marie Brouwer, and Jan BF van Erp. Combining and comparing eeg, peripheral physiology and eye-related measures for the assessment of mental workload. *Frontiers in neuroscience*, 8, 2014.

[11] Stephen H Fairclough. Fundamentals of physiological computing. *Interacting with computers*, 21(1-2):133–145, 2008.

[12] Scott Makeig, Tzyy-Ping Jung, and Terrence J Sejnowski. Awareness during drowsiness: Dynamics and electrophysiological correlates. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 54(4):266, 2000.

[13] Maarten AS Boksem, Theo F Meijman, and Monicque M Lorist. Effects of mental fatigue on attention: an erp study. *Cognitive brain research*, 25(1):107–116, 2005.

[14] Saroj KL Lal and Ashley Craig. A critical review of the psychophysiology of driver fatigue. *Biological psychology*, 55(3):173–194, 2001.

[15] Claudio Stampi, Polly Stone, and Akihiro Michimori. A new quantitative method for assessing sleepiness: the alpha attenuation test. *Work & Stress*, 9(2-3):368–376, 1995.

[16] Jan BF van Erp, Maarten A Hogervorst, and Ysbrand D van der Werf. Toward physiological indices of emotional state driving future ebook interactivity. *PeerJ computer science*, 2:e60, 2016.

[17] Elena Rogado, José Luis Garcia, Rafael Barea, Luis Miguel Bergasa, and Elena López. Driver fatigue detection system. In *Robotics and Biomimetics, 2008. ROBIO 2008. IEEE International Conference on*, pages 1105–1110. IEEE, 2009.

[18] Gianluca Borghini, Laura Astolfi, Giovanni Vecchiato, Donatella Mattia, and Fabio Babiloni. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews*, 44:58–75, 2014.

[19] Saroj KL Lal and Ashley Craig. Driver fatigue: electroencephalography and psychological assessment. *Psychophysiology*, 39(3):313–321, 2002.

[20] Chunlin Zhao, Min Zhao, Jianpin Liu, and Chongxun Zheng. Electroencephalogram and electrocardiograph assessment of mental fatigue in a driving simulator. *Accident Analysis & Prevention*, 45:83–90, 2012.

[21] Wen Chieh Liang, John Yuan, Deh Chuan Sun, and Ming Han Lin. Changes in physiological parameters induced by indoor simulated driving: Effect of lower body exercise at mid-term break. *Sensors*, 9(9):6913–6933, 2009.

[22] Peter TF Raggatt and Shirley A Morrissey. A field study of stress and fatigue in long-distance bus drivers. *Behavioral medicine*, 23(3):122–129, 1997.

[23] M Myrtek, E Deutschmann-Janicke, H Strohmaier, W Zimmermann, S Lawerenz, G Brügner, and W Müller. Physical, mental, emotional, and subjective workload components in train drivers. *Ergonomics*, 37(7):1195–1203, 1994.

[24] Ross J Apparies, Todd C Riniolo, and Stephen W Porges. A psychophysiological investigation of the effects of driving longer-combination vehicles. *Ergonomics*, 41(5):581–592, 1998.

[25] G Dorfman Furman, A Baharav, C Cahan, and S Akselrod. Early detection of falling asleep at the wheel: A heart rate variability approach. In *Computers in Cardiology, 2008*, pages 1109–1112. IEEE, 2008.

[26] Bruce Mehler, Bryan Reimer, and Joseph F Coughlin. Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task: an on-road study across three age groups. *Human factors*, 54(3):396–412, 2012.

[27] Bryan Reimer, Bruce Mehler, Joseph F Coughlin, Kathryn M Godfrey, and Chuanzhong Tan. An on-road assessment of the impact of cognitive workload on physiological arousal in young adult drivers. In *Proceedings of the 1st international conference on automotive user interfaces and interactive vehicular applications*, pages 115–118. ACM, 2009.

[28] Jennifer A Healey and Rosalind W Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems*, 6(2):156–166, 2005.

[29] Anne-Marie Brouwer, Chris Dijksterhuis, and Jan BF van Erp. Physiological correlates of mental effort as manipulated through lane width during simulated driving. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 42–48. IEEE, 2015.

[30] Chris Dijksterhuis, Karel A Brookhuis, and Dick De Waard. Effects of steering demand on lane keeping behaviour, self-reports, and physiology. a simulator study. *Accident Analysis & Prevention*, 43(3):1074–1081, 2011.

[31] Chris Dijksterhuis, Dick de Waard, Karel A Brookhuis, Ben LJM Mulder, and Ritske de Jong. Classifying visuomotor workload in a driving simulator using subject specific spatial brain patterns. *Frontiers in neuroscience*, 7, 2013.

[32] Melissa A Birkett. The trier social stress test protocol for inducing psychological stress. *JoVE (Journal of Visualized Experiments)*, (56):e3238–e3238, 2011.

[33] Mats Gillberg, Göran Kecklund, and Torbjörn Åkerstedt. Relations between performance and subjective ratings of sleepiness during a night awake. *Sleep: Journal of Sleep Research & Sleep Medicine*, 1994.

[34] Michael Ingre, TorbjÖRn ÅKerstedt, BjÖRn Peters, Anna Anund, and GÖRan Kecklund. Subjective sleepiness, simulated driving performance and blink duration: examining individual differences. *Journal of sleep research*, 15(1):47–53, 2006.

[35] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, 23(10):1175–1191, 2001.

[36] Michael E Dawson, Anne M Schell, and Diane L Filion. The electrodermal system. *Handbook of psychophysiology*, 2:200–223, 2007.

[37] Marek Malik. Heart rate variability. *Annals of Noninvasive Electrocardiology*, 1(2):151–181, 1996.

[38] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[39] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[40] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[41] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[42] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[43] Pouya Bashivan, Irina Rish, Mohammed Yeasin, and Noel Codella. Learning representations from eeg with deep recurrent-convolutional neural networks. *arXiv preprint arXiv:1511.06448*, 2015.

[44] Zhi-Hua Zhou and Xu-Ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):63–77, 2006.

[45] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[46] Ivan Tomek. Two modifications of cnn. *IEEE Trans. Systems, Man and Cybernetics*, 6:769–772, 1976.

[47] Alexandre Dalyac, Murray Shanahan, and Jack Kelly. Tackling class imbalance with deep convolutional neural networks. *Imperial College*, pages 30–35, 2014.

[48] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, volume 9, pages 249–256, 2010.

[49] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629*, 2016.

[50] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[51] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.