**Deep Residual Networks for Detecting Driver Stress from PPG signal**

**Abstract**

**The vast majority of methods for driver stress detection rely on complex multimodal signals that require intrusive and expensive methods of acquisition. In the presented approach we hypothesize that the PPG signal provides enough physiological information to be used as independent reliable stress indicators. To achieve this, we propose a novel end-to-end deep learning model that utilizes the Residual neural blocks empowered by multi-head mechanisms. By introducing input network layers which prioritize reducing spatial resolution and extracting key features, the proposed model eliminates manual feature engineering over the input PPG signal and contributes significantly to the overall efficiency of the model. We provide a comprehensive evaluation of the models generalization capability, network structure, and classification accuracy. Experimental results demonstrate that our approach achieves superior performance compared to other state-of-the-art methods, while minimizing system complexity and cost. Therefore, it provides high potential for integration into real-world driving experience.**

**Introduction**

Detection of driver stress is a critical aspect of automotive safety and human-machine interaction research. While several sensor modalities such as Electrodermal Activity (EDA), skin temperature, or electrocardiogram (ECG) might be employed to monitor driver stress [XXX], this paper focuses on the utilization of Photoplethysmography (PPG) as the sole modality for stress detection.

The PPG signal serves as a valuable indicator of dynamic blood volume pulse (BVP) in the body’s extremities, providing insight into various physiological and cardiovascular parameters, including heart rate (HR), heart rate variability (HRV), oxygen saturation (SpO2) and respiratory rate [1]. Notably, the connection between the heart and the brain through the autonomic nervous system (ANS) means that changes in emotional states can influence heart rate and blood ejection rate. During stressful situations, the sympathetic part of the ANS is activated, leading to a rapid increase in heart activity. This accelerates blood pumping, ensuring a rapid oxygen delivery to the body’s organs in order to enable the individual to react quickly. Increased respiratory rate causes changes to the shape of PPG signal which makes it a valuable parameter for stress assessment. In stressful situations, HR and HRV fluctuations in ANS activity lead to alterations in HR and HRV patterns, rendering them effective indicators of a driver’s stress level [2]. Both of these features can be extracted from the PPG signal.

While multi-modal sensor setups could be valuable in certain applications, there are compelling motivations to rely solely on PPG sensors for driver stress detection in our research. They are listed below:

* Non-intrusiveness: PPG is a non-invasive optical technique that measures changes in blood volume through skin using a relatively simple sensor. It does not require direct contact with the skin or electrodes, making it less intrusive and avoiding potential irritation.
* Comfort and user acceptance: PPG can be seamlessly integrated into common wearable devices like smartwatches or fitness trackers. This makes it more acceptable and comfortable for drivers, as they can continue their regular activities without additional discomfort, unlike other cumbersome sensors.
* Unobtrusive continuous monitoring: PPG allows for continuous monitoring of physiological parameters without the need for frequent recalibration or sensor readjustment. In contrast, EDA sensors may require periodic adjustment to maintain proper skin contact, which can be impractical while driving.
* Rich physiological information: PPG provides valuable physiological data correlated to heart rate, heart rate variability, and arterial stiffness, which can be indicative of stress levels. These parameters are often used as reliable stress indicators, eliminating the need for additional sensors like skin temperature sensors.

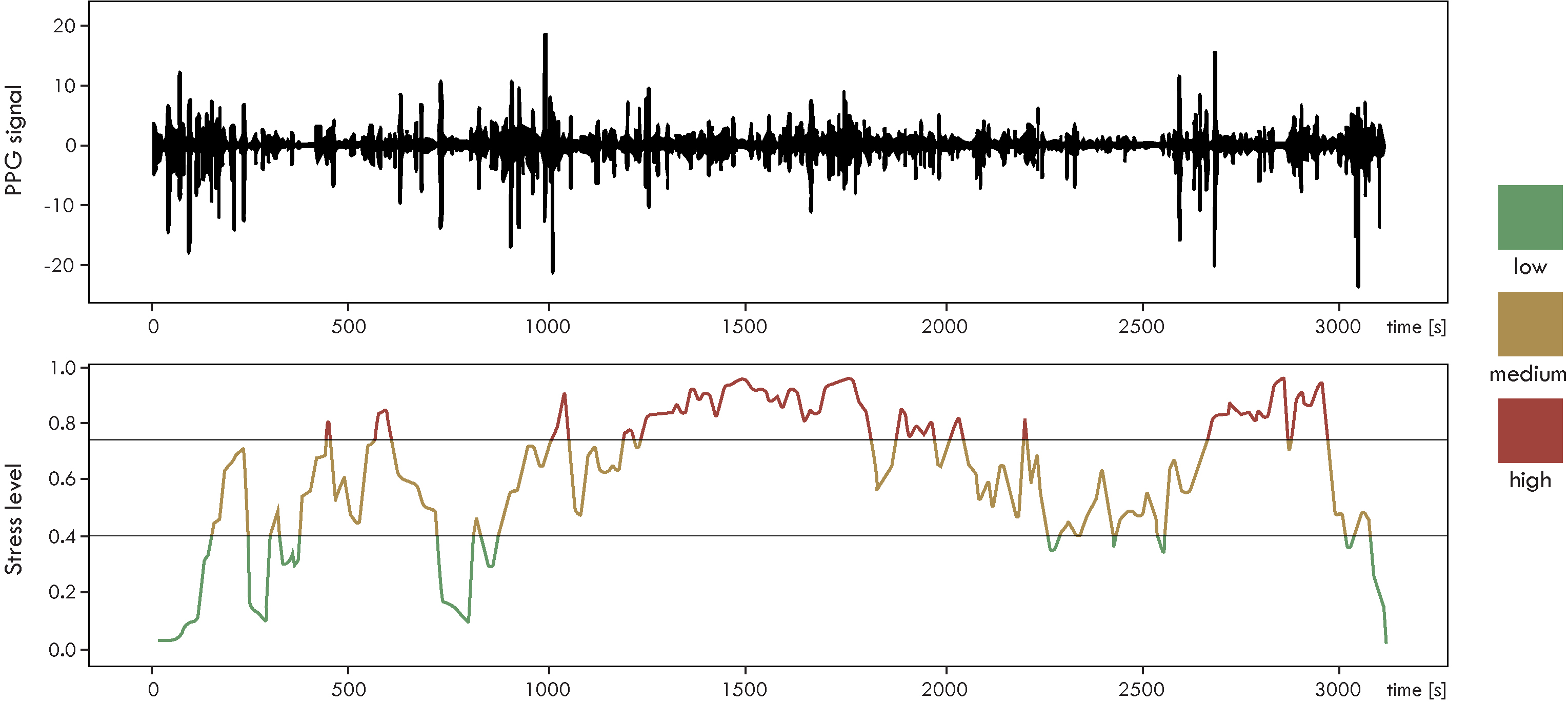


Figure 1. Example of PPG signal (top) and estimated stress score (bottom) during a drive

* Wearability and integration: PPG can be seamlessly integrated into various wearable devices that are increasingly popular among drivers. This integration allows for continuous, unobtrusive data collection, facilitating real-time stress detection without the need for additional sensors.
* Cost-efficiency: By relying solely on PPG, the cost of sensor deployment and maintenance can be significantly reduced compared to multi-modal sensor setups, making it a cost-efficient option for stress detection in automotive applications.

Previous advantages make PPG sensors a very promising choice for developing efficient and user-friendly driver stress detection systems, contributing to enhanced road safety and driver well-being. Figure 1 illustrates the temporal dynamics of BVP (top graph) and stress level (bottom graph).

In the context of driver stress detection from PPG signals, after splitting the input data into sliding windows, our aim is to classify them into several categories of stress level. Examples of sliding windows belonging to three categories of stress (low, medium, and high) are shown in figure Figure 2.

From the previous figures it is evident that the PPG signal exhibits a significant level of noise that might be an insurmountable challenge for traditional machine learning techniques, where manual feature extraction has to be performed.

However, recent advances in Deep Learning which has made a significant progress in Computer Vision [REF image net] and Natural language Processing [REF], has led researchers to explore its application for solving complex time-series classification problems. One of the strengths of deep learning methods is its capacity to automatically extract  novel discriminative features from the low-level noisy input signal.

By empowering the PPG physiological signal with modern Deep Learning algorithms for time-series data analyses, we have developed a stress detection method that could be seamlessly integrated into the driving experience to provide accurate monitoring of driver stress levels while minimizing system complexity and cost. To achieve this, we propose a novel end-to-end deep learning model that utilizes the residual neural blocks empowered by multi-head mechanisms. The proposed model completely eliminates manual feature engineering, by introducing carefully engineered input network layers which prioritize reducing spatial resolution and extracting key features. A comprehensive evaluation of the models generalization capability, network structure, and classification accuracy, demonstrate that our approach achieves superior performance compared to other state-of-the-art methods, while minimizing system complexity and cost. Our paper makes two main contributions: 1) Presenting a highly effective deep learning  model for driver stress detection relying solely on PPG signal, 2) Demonstrating that the relatively simple deep learning architecture, based on ResNet blocks combined with multi-head approach, could achieve superior performance, thus having potential to be integrated into real-world automotive applications.

The rest of the paper is organized in XXX main sections. Section II introduces...Section III describes PPG-based methods... In section IV, we discuss emerging... Finally, the conclusion summarizes the paper proposal and briefly..

.

**Related work**

*Traditional ML methods vs Deep Learning methods for effective time-series classification?*

Detecting stress levels from a PPG signal involves addressing a time series problem. In the context of data analysis, the objective is to recognize meaningful patterns, trends, and fluctuations within the collection of PPG data points recorded at regular time intervals. We approach driver stress identification as a classification problem over PPG time series data. Therefore, it demands robust Machine Learning (ML) models capable of capturing intricate temporal dependencies and patterns within PPG physiological data.

Earlier time-series classification approaches relied on traditional ML methods, such as Decision Trees, Linear Regression, and Support Vector Machines (SVM). These methods typically necessitated manual feature engineering, which involves creating new data attributes from existing ones to provide more relevant information to the ML model. Feature extraction can occur across various domains using linear or non-linear methods. This is done by transforming, combining, or aggregating existing features. The primary goal is often to reduce the number of features while retaining important information.

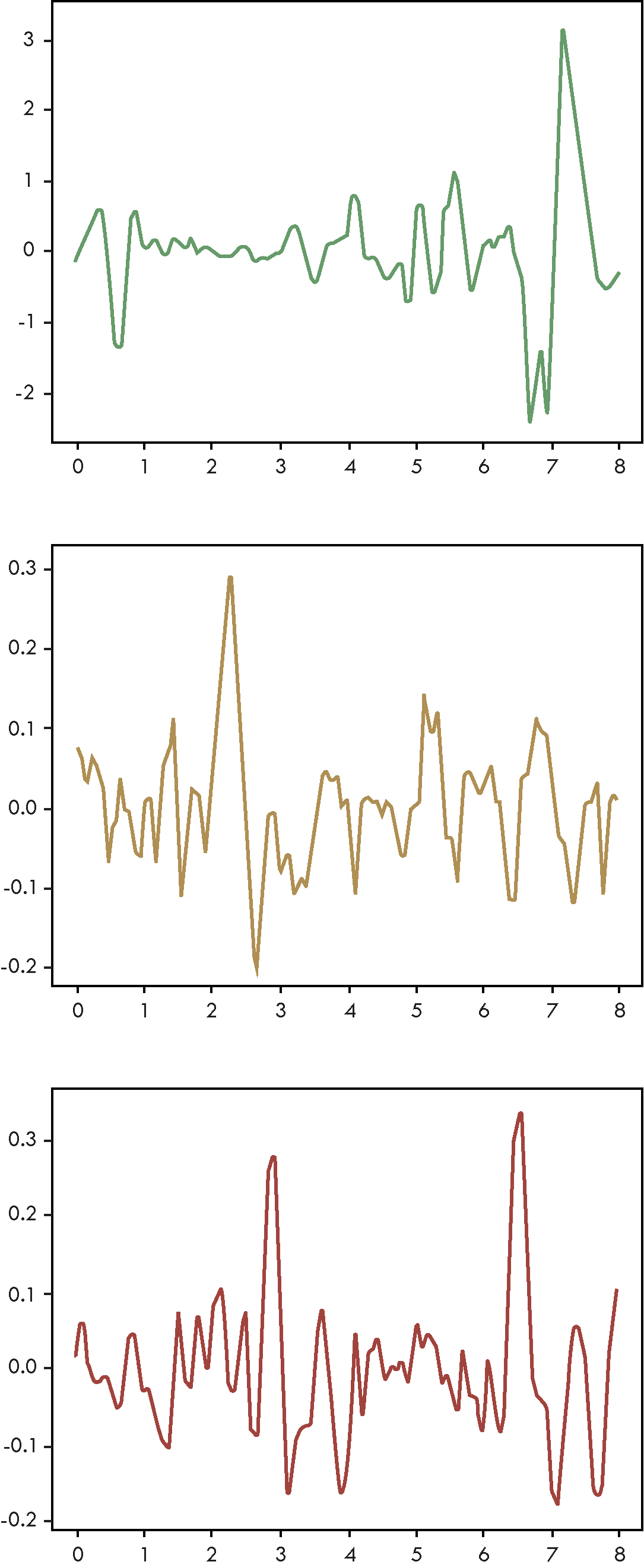


Figure 2. Examples of sliding windows during low (a), medium (b) and high (c) stress intervals

Deep Learning presents an alternative approach in which models can autonomously learn features from raw time series data, eliminating the need for manual feature engineering. Convolutional and recurrent layers are commonly employed for this purpose.

While traditional ML models generally tend to be simpler with fewer parameters, this simplicity can hinder their ability to generalize effectively when dealing with noisy or complex data, thus leading to reduced accuracy of results.  Modern Deep Learning models can be highly complex, featuring numerous layers and parameters. This complexity enables them to capture intricate data patterns but also renders them susceptible to overfitting if not properly regularized.

A notable limitation of Deep Learning methods is its higher computational requirements, particularly with deep architectures and large input datasets. Furthermore, interpreting and explaining complex Deep Learning models can be challenging, while traditional ML models often offer greater interpretability due to the transparency of the feature engineering process. This was a strong motivation for our research, to investigate optimal design of a deep learning model with limited complexity and high accuracy for time-series classification. Two prominent neural network architectures have been mostly investigated for time-series analyses: Long Short-Term Memory (LSTM) and Residual Networks (ResNet). Both architectures represent formidable candidates for effective PPG time series analysis in the context of driver stress detection. We will consider their respective advantages and limitations.

**LSTM for PPG time-series data analyses**

* a type of recurrent neural network (RNN), is inherently designed to capture and model long-range temporal dependencies within sequential data. In the context of driver stress detection, this is crucial as it allows the model to consider past physiological signals (e.g., PPG data) when making predictions, effectively learning patterns and trends over time.
* Memory Cells: LSTM includes memory cells that can capture information over extended time steps, mitigating the vanishing gradient problem often encountered in traditional RNNs. This enables it to maintain context and effectively model temporal relationships.
* Handling Variable Sequence Lengths: LSTMs can handle time series data with varying lengths, a characteristic often encountered in real-world scenarios. This adaptability is particularly relevant when dealing with driver stress detection, where stress-inducing events may occur at irregular intervals.
* Robustness to Noise: LSTMs are inherently robust to noisy data and can filter out irrelevant information, making them well-suited for handling noisy physiological signals that may be present in the driving environment.

Limitations of LSTM for PPG data analysis:

* Computational Complexity: LSTM networks can be computationally expensive with a large number of parameters, which may require substantial computational resources for training and inference. This complexity may pose challenges in real-time applications or when dealing with large datasets.
* Slower Training: LSTMs and RNNs, in general, process data sequentially, one time step at time. This creates a dependency on the previous time step for each subsequent one.
* Risk of Overfitting: LSTMs are prone to overfitting, especially when dealing with small datasets. Careful regularization, validation and hyperparameter tuning are often necessary to mitigate this risk and achieve optimal performance.

**ResNet for PPG time-series data analyses**

Advantages of ResNet:

* Feature Extraction: ResNets excel at feature extraction, allowing them to automatically learn and represent important hierarchical features within the time series data. This can be highly beneficial when dealing with complex physiological PPG signals as it reduces the need for hand-crafted feature engineering.
* Residual Connections: The introduction of residual connections in ResNet architectures mitigates the vanishing gradient problem, facilitating the stable and efficient training even in very deep networks. This is advantageous when dealing with intricate patterns in time series PPG data.
* Parallel Processing: ResNets can process multiple temporal resolutions in parallel, which can be advantageous in capturing both short-term and long-term patterns within the same model.

Limitations of ResNet:

* Limited Sequential Context: While ResNet can capture local temporal dependencies, it may not be effective enough in modeling long-range sequential relationships. However, when dealing with the PPG signal in the context of detecting driver stress, it's essential to note that these temporal dependencies generally do not extend over extended intervals. This is because the PPG signal typically exhibits a repetitive, quasi-periodic pattern associated with each heartbeat. The complete waveform for a single heartbeat impulse is relatively brief, typically lasting only around a second. Additionally, PPG sensors in wearable devices often have lower sampling frequencies, usually within the range of 50 to 200 Hz.
* Potentially Larger Model Sizes: Due to their depth, ResNets may have more parameters when compared to LSTMs. This can result in larger model sizes, potentially affecting memory requirements. While this can impact memory requirements, it should not be a relevant limiting factor in modern automotive hardware.

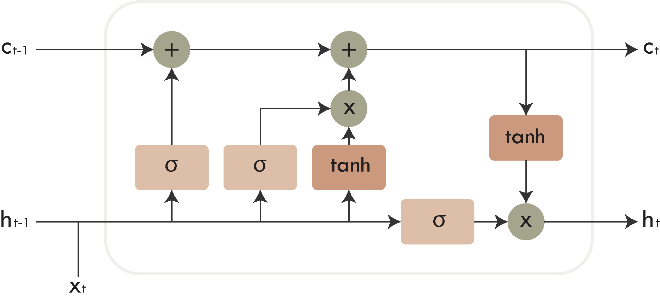


Figure 3. LSTM cell

The concrete choice between LSTM and ResNet depends on the specific characteristics of the signal, the computational resources available, and the desired trade-off between modeling temporal dependencies and feature extraction. While LSTMs excel at capturing long-term sequential dependencies, ResNets deep architecture and their efficient training offer powerful feature extraction capabilities for modeling complex patterns.

**Method**

**Architectures – LSTM, ResNet**

When compared to traditional machine learning approaches, neural networks offer the advantage of automatic feature extraction, eliminating the need for manually extracting and choosing features. In the context of time-series problems, Recurrent Neural Networks (RNNs) are one of the preferred options because they enable parameter sharing, rendering them suitable for handling sequences of varying lengths. A closely related concept involves the application of 1D convolution, often referred to as temporal convolution, as this operation also enables the sharing of parameters across different time points within the data.

The idea behind RNNs lies in the presence of a hidden state within each unit, which can be regarded as the memory of the unit. As a result, each recurrent layer receives two inputs: a vector of outputs from the preceding layer and the vector of states from the same layer at the previous time step. For these networks, computing the gradient of the loss function with respect to the parameters can be an expensive operation which is sequential in nature because each time step may only be computed after the previous one. This process is commonly referred to as back-propagation through time.

Learning long-term dependencies with RNNs can be challenging because gradients propagated across many stages tend to vanish or, less frequently, explode. Even if the recurrent network would remain stable, the challenge of handling long-term dependencies persists due to exponentially smaller weights (characterized by the multiplication of numerous Jacobians) of long-term interactions compared to short-term ones. This implies that features from the beginning of sequences tend to be “forgotten”. To address these challenges, gated RNNs are employed. They are based on the idea of establishing paths through time with derivatives that neither vanish nor explode. One of the most commonly used types of gated RNNs are LSTM networks.

The LSTM layers within neural networks consist of a set of gates which regulate the flow of information. Each module comprises a forget gate, an input gate, an output gate, and a cell state. This is illustrated in Figure 3.

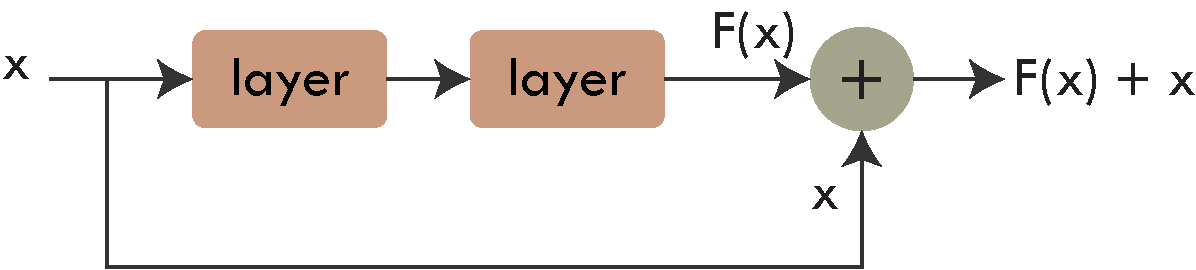


Figure 4. Residual block

The role of the forget gate is to determine the extent to which information from the previous cell should be preserved. This is achieved using the sigmoid function, whose output ranges from 0 (indicating the decision to forget everything) to 1 (indicating the decision to remember everything).

For a specific time step and a given cell , this operation can be presented in the following way:

Here, represents the current input vector, is the current hidden layer vector, and , and correspond to the biases, input weights and recurrent weights for the forget gates, respectively.

The LSTM cell internal state is updated in the following way:

The ouput of the input gate is computed in a similar manner to the forget gate:

Finally, the output of the LSTM cell is controlled by the output gate:

The input gate serves to regulate the amount of new information to be incorporated into the cell state and the output gate determines what information should be presented as the LSTM cell’s output.

Our approach involves the utilization of Residual Networks, initially introduced in a 2015 paper [3]. While it’s well-recognized that deeper networks possess the capability to learn more complex features, a significant challenge arises in the form of vanishing and exploding gradients. This challenge can be partially mitigated through the incorporation of initial normalization and intermediate normalization layers.

ResNets were designed to address this challenge by enabling increased network depth without compromising performance. This is achieved by having the neural network learn a residual mapping denoted as F(x) = H(x) – x, rather than the desired mapping H(x). This transformation reframes the original mapping into F(x) + x.

The realization of F(x) + x is done through the use of “shortcut connections” in a feedforward neural network, as illustrated in Figure 4. These connections are able to bypass one or more layers, introducing neither additional parameters nor computational complexity.

The underlying hypothesis suggests that learning such a mapping is more manageable. Even if the identity mapping were considered optimal, it would be more feasible to drive the residual component toward zero than to precisely identify an identity mapping.

**Proposed Deep Neural Network Architectures**

We designed and implemented two distinct architectures, one with Residual Connections and another with LSTM layers. Our primary emphasis was on the Residual Networks, where we further investigated several modifications to improve the performance.

The input structure for all architectures follows the same pattern. Much like in the original ResNet structure, we begin with a Max Pooling layer, which serves a dual purpose. Firstly, it contributes to the reduction of dimensionality within the input data. Secondly, it helps with feature selection. By selecting the maximum value from each of the pooling regions, the network effectively retains most of the prominent features, filtering out less relevant and noisy information.

Moving forward, we introduce a strided convolution layer, a component typically used for downsampling, akin to pooling layers [4]. As an extra layer in our design, we also incorporate an Average Pooling layer. This layer utilizes a filter size of 4x1 and a stride of 4, leading to notable simplification of the input vector.

In essence, our input structure prioritizes two key aspects: reducing spatial resolution and selecting key features. This contributes significantly to enhancing the overall efficiency of the model.

The output structure of both layers is comprised of a Dropout layer, a Fully Connected layer and a Softmax layer. For the Dropout layer, we employed a dropout probability of 0.2. This layer, along with weight decay, plays a vital role in regularization. Such techniques hold significance because neural networks sometimes tend to capture statistical noise present in the dataset, which can potentially result in overfitting.

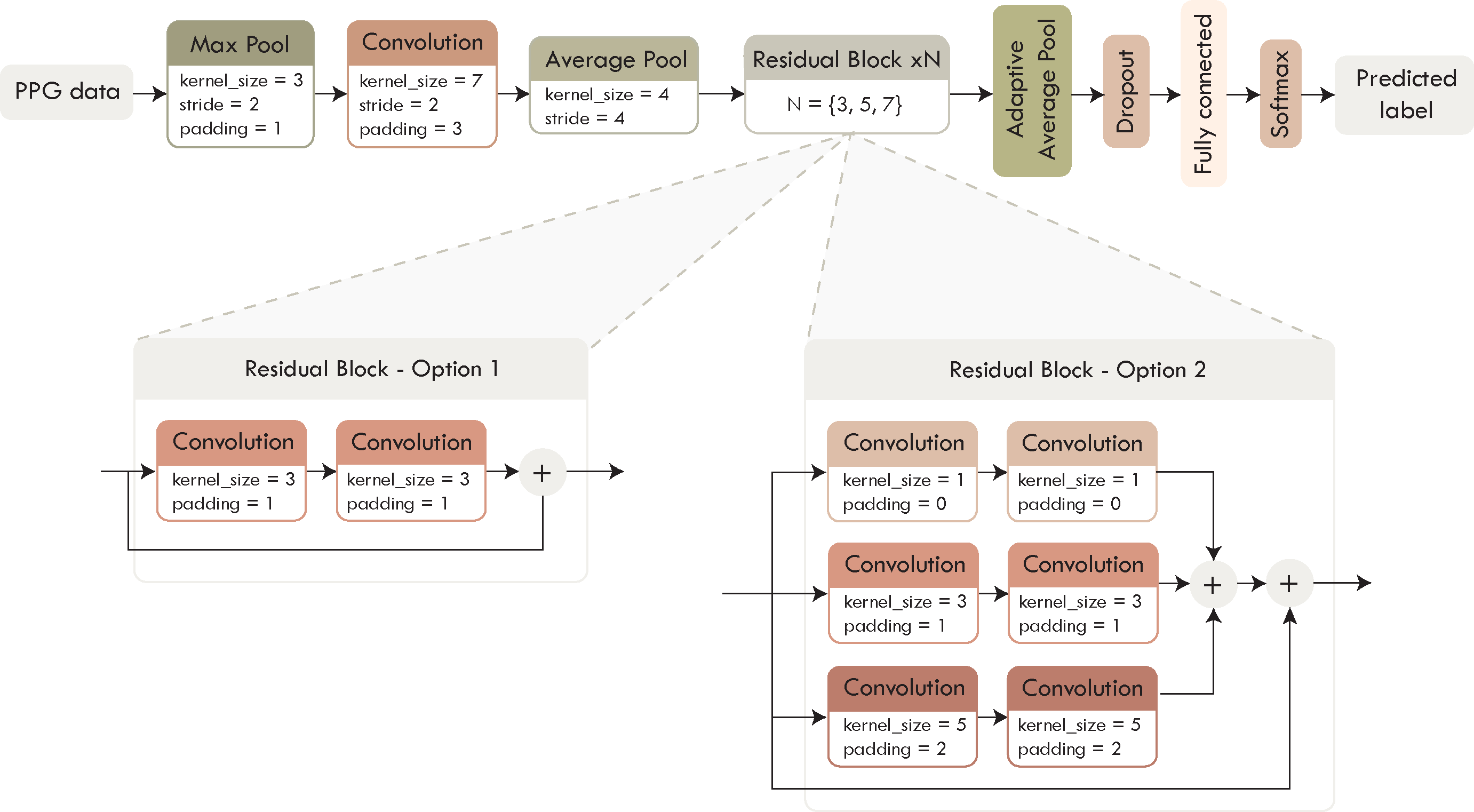


Figure 5. ResNet architecture

Moving forward, the Fully Connected layer is used for processing and combining the features extracted in earlier stages of the network. These layers allow the network to learn complex, non-local patterns and relationships within the data. In the context of classification tasks, we've incorporated a Softmax layer, which is commonly used in neural networks to transform the model's output into probability distributions over multiple classes. This enables us to make predictions with confidence scores for each class.

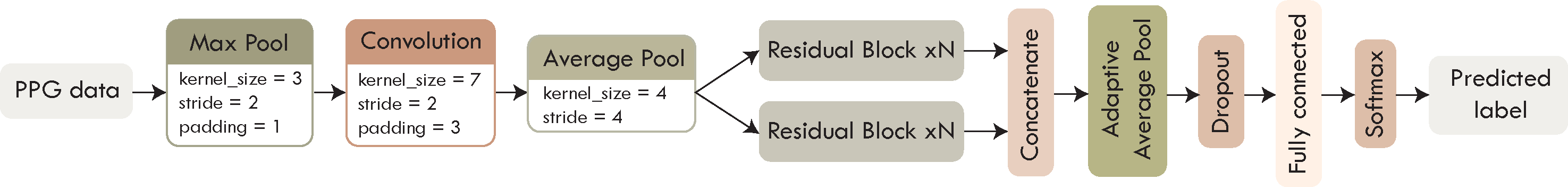


Figure 6. Multi-head ResNet architecture

In the ResNet-based architecture, following the initial input processing, we introduce a series of N Residual Blocks. These blocks can be stacked, and we evaluated the performance for different values of N, specifically N = 3, 5, and 7.

As is shown in Figure 5, there are two options we considered for these Residual Blocks. The first one is the approach used in the original paper [3]. Here, each of these blocks consists of two convolutional layers and a shortcut connection. Regarding the choice of filter size for the convolutional layers, we’ve opted for a 3x1 configuration. While larger filter sizes have the potential to capture more global features from the input data, they also increase the number of parameters. This can slow down training and render the model susceptible to overfitting. To ensure compatibility for the addition operation between the output and the shortcut, the padding is set to 1.

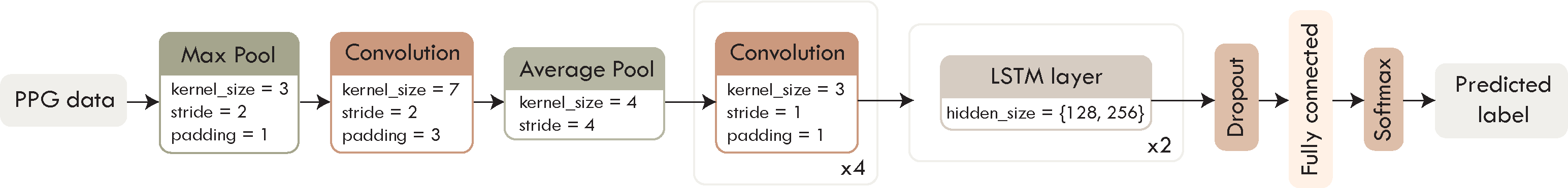


Figure 7. LSTM architecture

The second approach involves the use of multiple paths with different filter sizes (1x1, 3x1, 5x1), enabling each path to capture distinct information. The outputs from these paths are aggregated, and the result is combined with the shortcut.

Each of the convolutional layers comes with Batch Normalization, a method that enhances training stability. As previously mentioned, these layers play a crucial role in mitigating challenges such as exploding and vanishing gradients, as well as the potential for the network to get trapped in unfavorable local minima [5]. This is especially important when training very deep networks, as the variables in intermediate layers may take values with substantially varying magnitudes. Another benefit of utilizing Batch Normalization is that it reduces the training time. Moreover, it has a minor influence on regularization, which contributes to reducing the risk of overfitting.

To introduce non-linearity and enable the network to learn complex patterns, we employ the Rectified Linear Unit (ReLU) activation function. This function retains positive values as they are and sets negative values to zero.

Before reaching the output layers, we incorporate an Adaptive Average Pooling layer that serves to reshape and streamline the data vector. Unlike traditional average pooling, which operates with fixed kernel sizes, the adaptive variant dynamically adjusts its kernel size based on the dimensions of the input data. This adaptability ensures that the output of the pooling layer is of a consistent size, regardless of the input's spatial dimensions.

As shown in Figure 6, we have also explored the option of having multiple branches of parallel layers which are concatenated before reaching the output structure. Such network configurations are often referred to as multi-head architectures [6], [7]. The underlying principle is that, even if the branches consist of identical layers, they can capture somewhat different features during the learning process.

Having multiple paths enables the increase of network complexity without increasing the depth. The efficiency of learning temporal dependencies in neural networks is significantly influenced by the length of the pathways traversed by signals moving forward and backward within the network. In essence, the shorter the paths between the input and output sequences, the more effectively the network can capture and learn long-range dependencies in the data [8].

The design of the LSTM-based network is rooted in the DeepConvLSTM framework [9], which is characterized by a sequence of four convolutional layers followed by two LSTM layers, as presented in Figure 7. Like in the previous architecture, it’s worth noting that each of the convolutional blocks is completed by Batch Normalization and ReLU activation.

The LSTM layers are arranged in a stacked configuration, where the output of one LSTM layer serves as the input to the next.   
Within each LSTM layer, there are M hidden units. To be precise, we assessed the model's performance for M values of 128 and 256. These hidden units are responsible for learning and representing complex patterns and temporal dependencies within the input data. This stacking of LSTM layers facilitates the modeling of sequential data by gradually building a deeper understanding of both short-term and long-term dependencies.

The input shape of the data can be denoted as *b* x *s* x *c*, where *b* corresponds to the mini-batch size, *s* signifies the number of samples contained within each sliding window (sampling rate multiplied by the time duration of the window), and *c* represents the number of data channels.

Given our exclusive utilization of BVP data, the channel count can either be 1 or 2 depending on whether the data source comprises a single wrist (in the case of the WESAD dataset [10]) or data from both wrists (in the case of the AffectiveROAD dataset [11]), as described in the following.

**Datasets for Evaluation**

For assessing the drivers’ stress level, we utilized two publicly available datasets. AffectiveROAD dataset [9], a publicly available resource which contains 13 driving sessions performed by 10 distinct drivers. Each drive spanned approximately 85 minutes, incorporating a 30-minute rest interval. These routes include a variety of road types and environmental conditions, leading to varying degrees of stress. A human experimenter, situated in the back seat, subjectively estimated the stress levels using a laptop-based slider scale that ranged from 0 (no stress) to 1 (extremely high stress). After the drives, the drivers were asked to review and, if needed, change the stress scores based on their perception.

Throughout the course of the experiment, participants wore two physiological monitoring devices: the Empatica E4 sensor on both wrists and the Zephyr Bioharness 3 on the chest. Using these sensors enabled the capture of many physiological signals, including Blood Volume Pulse (BVP), Electrodermal Activity (EDA), skin temperature (TEMP), and motion (ACC).

Additionally, the performance evaluation was conducted using the Multimodal Dataset for Wearable Stress and Affect Detection (WESAD) dataset [10], which is publicly available as well. The data was collected by using two recording devices: the RespiBAN Professional on the subject’s chest and the Empatica E4 on the wrist of the non-dominant hand. A total of 17 subjects participated in the study, but the data from two participants was excluded due to sensor malfunction.

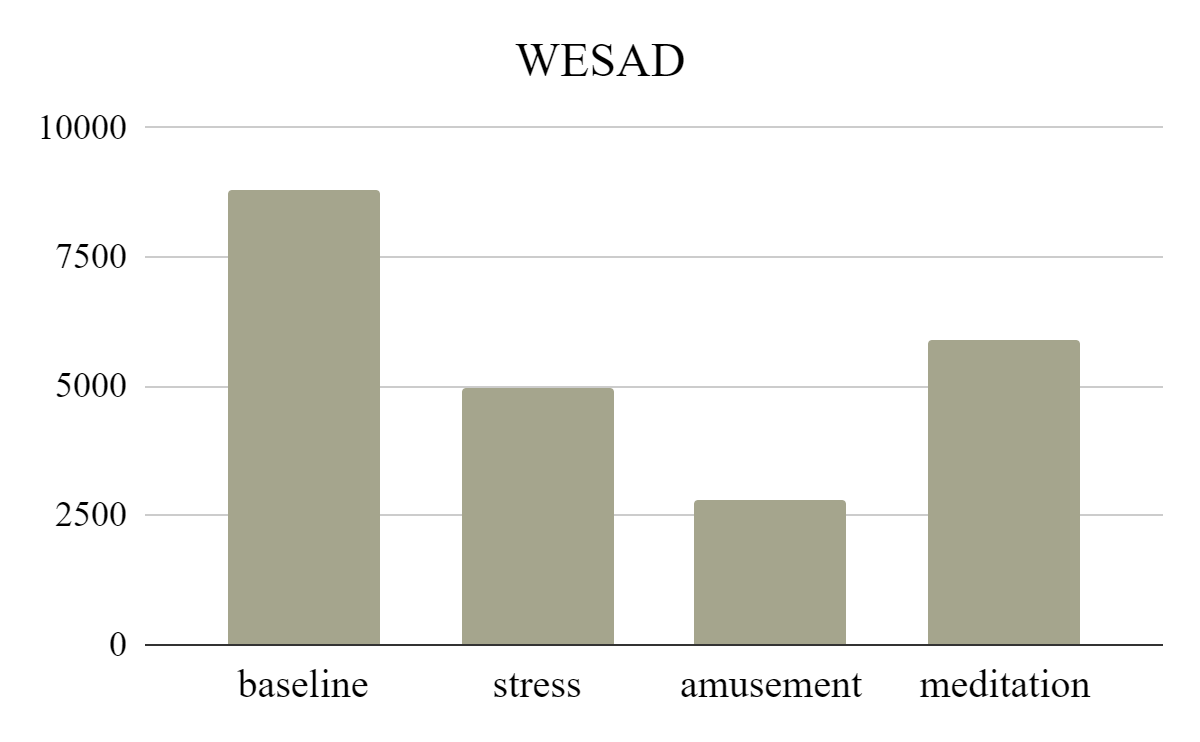
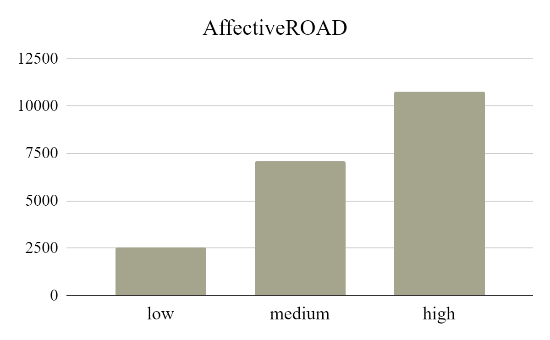


Figure 8. Class distribution for (a) AffectiveROAD and (b) WESAD datasets

To begin, the participants’ baseline condition was recorded for about 20 minutes. During this initial period, they were sitting or standing while reading materials such as magazines. Subsequently, they viewed 11 video clips, lasting a total of 392 seconds, with the intention of inducing amusement.

Following this, the subjects underwent the Trier Social Stress Test (TSST), which included delivering a speech in front of a three-person panel and performing a mental arithmetic task involving counting backward from 2023 to 0 in increments of 17. This phase, on average, spanned approximately 10 minutes. Following both the amusement and stress conditions, participants followed a guided meditation designed to help them return to the baseline state.

It is evident that both datasets utilize the E4 sensor which records BVP at a rate of 64 Hz. For the subsequent analyses, only the BVP data from the E4 sensor was utilized. To ensure alignment between the BVP data and the label frequencies, we performed downsampling from 700 Hz (for the WESAD dataset) and upsampling from 4 Hz (for the AffectiveROAD dataset).

**Imbalanced Datasets Preprocessing**

After importing, the features of the data are standardized by removing the mean and scaling to unit variance, according to the following formula:

Here, *Xstd* represents the standardized data, *X* is the original data while *μ* and *σ* are the mean and standard deviation of the original data, respectively.

Given the substantial amount of data captured by wearable sensors, a segmentation process involving sliding windows is employed before feeding the data into the model. This segmentation involves dividing the data into fixed-length windows, where subsequent windows will have a certain degree of overlap. As demonstrated later in this paper, the evaluation outcomes are influenced by two key factors: the length of the sequence and the overlap ratio.

Depending on the dataset in use, each sliding window is assigned a suitable label. The WESAD dataset already has labels for 4 different affective states (baseline, stress, amusement, meditation) for each time step. If all the labels inside a sliding window match, that is the resulting label, if they don’t that window will be discarded.

On the other hand, AffectiveROAD has relative stress scores, as previously discussed, which is why a different label assignment approach is used. For each window, we calculate the average of the stress values contained within it and categorize it into one of three levels: low (0-0.4), medium (0.4 – 0.75), or high (0.75 – 1). This was done in the same manner as in [12]. It is assumed that the stress levels will be high during city driving, medium during highway driving and low during rest periods.

As depicted in Figure 8 the distribution of labels across classes exhibits some degree of imbalance, potentially leading to model bias favoring the majority class. The values depicted in the graphs correspond to the scenario where 8-second windows with a 75% overlap were used. To mitigate this issue, various methods can be employed, such as oversampling and undersampling. However, given that the utilized datasets do not exhibit extreme imbalances, our primary strategy was to apply Class Weights.

This is a widely used technique that involves assigning distinct weights to classes based on their prevalence within the dataset. The objective is to enhance the significance of the minority class while diminishing the impact of the majority class during the model's training process. These class weights are seamlessly integrated into the model's loss function during training. Each class's weight is inversely proportional to its frequency in the dataset, resulting in higher weights being assigned to the minority classes.

These class weights can be calculated in the following manner:

Where *N* represents the total number of samples, *Ni* is the number of samples in class *i* and *k* is the total number of classes.

**Experimental Results**

**Initial comparison of ResNets and LSTMs**

In our evaluation process, we initially set the overlap between consecutive windows to 75%. To find the optimal window size, we evaluated the models’ performance across a range of window sizes, from 4 to 16 seconds. These results are summarized in Table 1, for the AffectiveROAD dataset and Table 2 for the WESAD dataset.

To maintain result consistency across multiple runs and to account for potential sources of variability, we fixed a constant seed value, thus mitigating the effects of dataset split randomness and other factors. For the initial results, the data was randomly divided into training and validation sets, with 80% allocated for training and the remaining 20% for validation.

Based on the results, we selected a window size of 8 seconds and analyzed the impact of changing the overlap ratio on the results. Specifically, we examined overlap ratios of 50%, 75% and 87.5% We chose these values because all the data within 50% overlap windows is also contained within 75% overlap windows, and similarly, 75% overlap windows are a subset of 87.5% overlap windows. As can be seen in Tables *3* and *4*, for both datasets and across all utilized architectures, it is evident that a higher overlap ratio consistently yields better results. A larger overlap ratio translates to a greater number of training samples.

Another notable observation is that in certain cases, the ResNet with 7 Residual Blocks performed less effectively than the one with 5 blocks. This highlights a common challenge encountered when increasing network depth: as neural networks become deeper, they become increasingly susceptible to overfitting. Overfitting implies that while the models excel in fitting the training data, they may struggle to generalize to new, unseen data [13]. This occurs because deeper networks have more parameters that can potentially capture noise within the training dataset.

When considering the LSTM architecture, in most cases, increasing the number of hidden units within LSTM layers leads to improved performance, but it also increases computational demands and time requirements.

This performance analysis highlights a notable superiority of ResNets over LSTM networks, with ResNets consistently delivering better results. It’s worth noting that training recurrent networks, such as LSTMs, tends to be considerably slower compared to CNNs. This is because RNNs maintain hidden states that need to be updated at each time step, and this operation is more computationally expensive than the simple convolutional operations performed by CNNs

For the AffectiveROAD dataset, we also investigated how performance metrics change when using data from only one of the wrists for the scenario of 8-second windows with 75% overlap. This comparison is presented in Table 5. As evident, the models demonstrate decreased performance when provided with less information. Interestingly, the utilization of data from the left wrist provides better results compared to the right wrist.

Table 1. AffectiveROAD dataset – comparison of different window sizes when the overlap ratios is fixed at 75%

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **AffectiveROAD** | 4s windows | | | | 8s windows | | | | 12s windows | | | | 16s windows | | | |
| F1 | Acc. | Prec. | Rec. | F1 | Acc. | Prec. | Rec. | F1 | Acc. | Prec. | Rec. | F1 | Acc. | Prec. | Rec. |
| ResNet (3 blocks) | 75.15 | 76.45 | 74.09 | 76.87 | 86.16 | 86.9 | 86.23 | 86.7 | 83.96 | 86.09 | 83.33 | 86.2 | 87.26 | 87.88 | 85.57 | 89.64 |
| ResNet (5 blocks) | 79.3 | 81.07 | 78.34 | 80.51 | 91 | 91.66 | 91.61 | 90.48 | 92.43 | 93.04 | 93.41 | 91.57 | 92.16 | 92.25 | 93.43 | 91.1 |
| ResNet (7 blocks) | 81.2 | 82.99 | 81.33 | 81.11 | 90.75 | 91.61 | 90.08 | 91.48 | 90.73 | 91.46 | 90.35 | 91.19 | 93.93 | 93.82 | 93.58 | 94.52 |
| LSTM (128 units) | 68.78 | 70.01 | 68.3 | 69.91 | 77.33 | 78.73 | 75.92 | 79.26 | 66.12 | 66.14 | 66.42 | 66.32 | 73.05 | 75.32 | 72 | 74.39 |
| LSTM (256 units) | 67.21 | 69.59 | 66.21 | 70.93 | 75.48 | 77.38 | 74.92 | 76.11 | 71.1 | 74.2 | 70.23 | 74 | 60.03 | 62.61 | 58.84 | 63.21 |

Table 2. WESAD dataset - comparison of different window sizes when the overlap ratios is fixed at 75%

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **WESAD** | 4s windows | | | | 8s windows | | | | 12s windows | | | | 16s windows | | | |
| F1 | Acc. | Prec. | Rec. | F1 | Acc. | Prec. | Rec. | F1 | Acc. | Prec. | Rec. | F1 | Acc. | Prec. | Rec. |
| ResNet (3 blocks) | 85.97 | 86.75 | 85.72 | 86.48 | 90.31 | 90.7 | 90.14 | 90.3 | 92.9 | 93.09 | 94.05 | 91.98 | 93.16 | 96.55 | 92.82 | 93.58 |
| ResNet (5 blocks) | 89.5 | 90.07 | 89.4 | 89.68 | 95.31 | 95.82 | 94.86 | 95.81 | 96.63 | 97 | 96.12 | 97.22 | 95.64 | 96.08 | 95.57 | 95.77 |
| ResNet (7 blocks) | 88.78 | 89.16 | 88.4 | 89.27 | 95.84 | 96.04 | 95.6 | 96.12 | 97.45 | 97.46 | 97.41 | 97.52 | 97.51 | 97.64 | 97.06 | 98.01 |
| LSTM (128 units) | 84.93 | 85.91 | 84.28 | 85.82 | 88.02 | 88.9 | 87 | 89.54 | 87.36 | 88.76 | 87.14 | 87.62 | 77.22 | 73.8 | 71.93 | 74.63 |
| LSTM (256 units) | 83.87 | 84.76 | 83.54 | 84.49 | 88.56 | 89.76 | 88.48 | 88.72 | 88.62 | 89.66 | 88 | 89.59 | 85.82 | 86.92 | 85.12 | 86.72 |

Table 3. AffectiveROAD dataset - effects of modifying window size when windows size is fixed at 8 seconds

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **AffectiveROAD** | 50% overlap | | | | 75% overlap | | | | 87.5% overlap | | | |
| F1 | Acc. | Prec. | Rec. | F1 | Acc. | Prec. | Rec. | F1 | Acc. | Prec. | Rec. |
| ResNet (3 blocks) | 66.72 | 70.61 | 65.88 | 67.86 | 86.16 | 86.9 | 86.23 | 86.7 | 94.81 | 95.03 | 94.74 | 94.88 |
| ResNet (5 blocks) | 70.85 | 74.68 | 72.37 | 69.71 | 91 | 91.66 | 91.61 | 90.48 | 96.55 | 96.86 | 96.27 | 96.84 |
| ResNet (7 blocks) | 69.67 | 74.63 | 74.82 | 67.64 | 90.75 | 91.61 | 90.08 | 91.48 | 97.33 | 97.47 | 96.74 | 97.97 |
| LSTM (128 units) | 51.21 | 55.89 | 52.52 | 52.43 | 77.33 | 78.73 | 75.92 | 79.26 | 90.39 | 90.85 | 89.57 | 91.36 |
| LSTM (256 units) | 53.92 | 58.39 | 52.92 | 53.85 | 75.48 | 77.38 | 74.92 | 76.11 | 90.87 | 91.36 | 89.83 | 92.06 |

Table 4. WESAD dataset - effects of modifying window size when windows size is fixed at 8 seconds

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **WESAD** | 50% overlap | | | | 75% overlap | | | | 87.5% overlap | | | |
| F1 | Acc. | Prec. | Rec. | F1 | Acc. | Prec. | Rec. | F1 | Acc. | Prec. | Rec. |
| ResNet (3 blocks) | 75.79 | 77.49 | 76.65 | 78.18 | 90.31 | 90.7 | 90.14 | 90.3 | 98.15 | 98.24 | 97.96 | 98.36 |
| ResNet (5 blocks) | 84.32 | 85.54 | 83.51 | 85.58 | 95.31 | 95.82 | 94.86 | 95.81 | 99.2 | 99.18 | 99.15 | 99.26 |
| ResNet (7 blocks) | 83.42 | 84.43 | 83.79 | 83.59 | 95.84 | 96.04 | 95.6 | 96.12 | 99.19 | 99.24 | 99.07 | 99.31 |
| LSTM (128 units) | 74.6 | 76.66 | 74.53 | 77.65 | 88.02 | 88.9 | 87 | 89.54 | 96.8 | 97.02 | 97.01 | 97.01 |
| LSTM (256 units) | 75.44 | 77.71 | 75.07 | 76.07 | 88.56 | 89.76 | 88.48 | 88.72 | 97.5 | 97.71 | 97.36 | 97.65 |

Table 5. AffectiveROAD dataset - results when using BVP data from one wrist only (8 second windows with 75% overlap)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **AffectiveROAD** | Left wrist | | | | Right wrist | | | |
| F1 | Acc. | Prec. | Rec. | F1 | Acc. | Prec. | Rec. |
| ResNet (3 blocks) | 79.68 | 81.26 | 77.82 | 82.64 | 77.55 | 79.29 | 76.64 | 78.92 |
| ResNet (5 blocks) | 86.72 | 88.13 | 88.11 | 85.59 | 81.79 | 83.29 | 81.87 | 82.4 |
| ResNet (7 blocks) | 87.48 | 88.37 | 87.78 | 87.29 | 85.2 | 86.19 | 84.36 | 86.45 |
| LSTM (128 units) | 61.07 | 62.22 | 59.76 | 67.24 | 60.54 | 62.88 | 59.13 | 63.88 |
| LSTM (256 units) | 65.25 | 68.08 | 64.25 | 68.57 | 59.53 | 63 | 58.72 | 61.29 |

**Performance of modified ResNets**

Based on the previously presented results, we determined that the ResNet architecture with 5 Residual Blocks delivered the optimal performance, surpassing the model with 3 blocks and often achieving results on par with, or even better than, the model with 7 blocks. In our subsequent evaluation, we explored variations of this network, which involved altering the Residual Blocks or introducing multiple branches of these blocks.

For this part of the evaluation, we employed 5-fold cross-validation instead of a conventional test-validation split. This approach splits the data into five subsets and iteratively trains and validates the model using different combinations of these subsets, leading to more reliable results.

Our primary focus centered on the 8-second windows with 87.5% overlap. However, we also conducted an evaluation for 75% overlap, as it tends to reveal more pronounced differences in the results.

The ideas behind the modifications were detailed earlier. Regarding the Residual Blocks, we explored different options with multiple paths employing various filter sizes: (1, 3, 5), (3, 5, 7) and (3, 5). As for the multi-head approach, we experimented with two and three heads, utilizing regular Residual Blocks. Additionally, we explored a configuration with two heads, specifically applying the (1, 3, 5) filter sizes, as this block type appeared to yield the most promising results.

**Comparative Evaluation**

To assess the performance of our models, we conducted comparisons with several prominent works, specifically considering the achieved accuracies. Although various approaches have been utilized for both of the test datasets [REF], our primary focus was on the ones utilizing Deep Neural Networks or exclusively relying on the BVP signal.

Starting with the AffectiveROAD dataset, the Deep Learning architecture employed in [19] consist of four convolution layers with varying numbers of filters (8, 32, 64, and 128) followed by a Global Average Pooling layer and a FC layer. The results were further enhanced by adding an LSTM layer after the FC layer. Here, all the E4 and BioHarness data was used.

Evaluation for both regression and classification was conducted in [20]. For the classification task, the data was divided into two categories: baseline and driving state. Feature extraction was performed, and the Random Forest algorithm was utilized for the classification task.

A non-neural network approach was used in [21] with the goal of differentiating between low (L) and high (H) stress classes. The drives were grouped into different profiles based on the driver’s physiological responses. Drive descriptors were created by calculating the mean of the features for all instances labeled as H. For clustering into different profiles, normalized spectral clustering was used. The signals from which the features were extracted were EDA and HR.

Moving on to the WESAD dataset, one of the methods employed a Self-Supervised Learning approach [14]. Here, an encoder was trained using distinct representations for the following signals: ACC, BVP, EDA and TEMP. The architecture was inspired by Google’s Inception [15], featuring blocks with multiple parallel layers whose outputs are stacked to form the final output.

Another neural network-based method [16] involved extracting features from BPV, EDA and ST signals, resulting in a total of 72 features. The architecture incorporated separate branches for each signal and a concatenating layer to learn shared encoded features. The overall loss was defined as the sum of losses from all branches. It’s worth noting that this approach used binary classification distinguishing between the stress state and non-stress states, which included baseline and amusement.

A high accuracy was achieved in [17]. This was done by utilizing the following signals: EDA, EMG, RESP and TEMP. Each of these signals was individually processed through a dedicated convolution block. The outputs from these blocks were then concatenated and passed through Fully Connected (FC) layers. Each of these blocks featured a convolution layer followed by a Max Pooling layer, and this sequence was repeated three times.

Lastly, in [18], three distinct classifiers were employed. When focusing solely on the BVP signal, the Random Forest classifier delivered the highest performance. Furthermore, this study delved into the impact of altering window sizes. The findings indicate that longer window sizes (up to 120 seconds) result in improved performance. Our choice for shorter window sizes was primarily motivated by the rapid changes that can occur in driving conditions, demanding a more responsive approach to stress detection.

Tables 6 and 7 provide a summary of the accuracies attained through the previously mentioned methods, alongside our own results.

Table 6. Comparison of our model with prominent works – AffectiveROAD dataset

|  |  |  |
| --- | --- | --- |
| Method | Accuracy | Utilized signals |
| SVM (rbf kernel) | 70 | EDA + HR |
| RBF (L2 regularization) | 83 |
| RF | 70.4 | BVP |
| NN | 87.95 | All E4 + BioHarness data |
| Proposed ResNet |  | BVP |

Table 7. Comparison of our model with prominent works - WESAD dataset

|  |  |  |
| --- | --- | --- |
| Method | Accuracy | Utilized signals |
| Self-supervised encoder network | 78.13 | ACC + BVP + EDA + TEMP |
| NN with 3 branches | 89.94 | BVP |
| RF | 81.4 | BVP |
| NN | 97.62 | ECG + EDA + EMG + RESP + TEMP |
| Proposed ResNet |  | BVP |

**Conclusion and discussion**

The study presented in this paper...

**Acknowledgement**

This research was supported by the Science Fund of the Republic of Serbia, project XAI4HEAT, grant no. XXXXXX.

**References**

[1] J. E. Sinex, “Pulse oximetry: Principles and limitations,” *Am. J. Emerg. Med.*, vol. 17, no. 1, pp. 59–66, Jan. 1999, doi: 10.1016/S0735-6757(99)90019-0.

[2] M. N. Rastgoo, B. Nakisa, A. Rakotonirainy, V. Chandran, and D. Tjondronegoro, “A Critical Review of Proactive Detection of Driver Stress Levels Based on Multimodal Measurements,” *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–35, Sep. 2019, doi: 10.1145/3186585.

[3] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[4] F. Bieder, R. Sandkühler, and P. C. Cattin, “Comparison of Methods Generalizing Max- and Average-Pooling,” 2021, doi: 10.48550/ARXIV.2103.01746.

[5] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.” arXiv, Mar. 02, 2015. Accessed: Oct. 24, 2023. [Online]. Available: http://arxiv.org/abs/1502.03167

[6] M. Canizo, I. Triguero, A. Conde, and E. Onieva, “Multi-head CNN–RNN for multi-time series anomaly detection: An industrial case study,” *Neurocomputing*, vol. 363, pp. 246–260, Oct. 2019, doi: 10.1016/j.neucom.2019.07.034.

[7] N. M. Foumani, L. Miller, C. W. Tan, G. I. Webb, G. Forestier, and M. Salehi, “Deep Learning for Time Series Classification and Extrinsic Regression: A Current Survey.” arXiv, Feb. 05, 2023. Accessed: Nov. 05, 2023. [Online]. Available: http://arxiv.org/abs/2302.02515

[8] A. Vaswani *et al.*, “Attention Is All You Need.” arXiv, Aug. 01, 2023. Accessed: Nov. 05, 2023. [Online]. Available: http://arxiv.org/abs/1706.03762

[9] F. Ordóñez and D. Roggen, “Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition,” *Sensors*, vol. 16, no. 1, p. 115, Jan. 2016, doi: 10.3390/s16010115.

[10] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, “Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection,” in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, Boulder CO USA: ACM, Oct. 2018, pp. 400–408. doi: 10.1145/3242969.3242985.

[11] N. E. Haouij, J.-M. Poggi, S. Sevestre-Ghalila, R. Ghozi, and M. Jaïdane, “AffectiveROAD system and database to assess driver’s attention,” in *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, Pau France: ACM, Apr. 2018, pp. 800–803. doi: 10.1145/3167132.3167395.

[12] C. Bustos, N. Elhaouij, A. Sole-Ribalta, J. Borge-Holthoefer, A. Lapedriza, and R. Picard, “Predicting Driver Self-Reported Stress by Analyzing the Road Scene,” in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, Nara, Japan: IEEE, Sep. 2021, pp. 1–8. doi: 10.1109/ACII52823.2021.9597438.

[13] L. Alzubaidi *et al.*, “Review of deep learning: concepts, CNN architectures, challenges, applications, future directions,” *J. Big Data*, vol. 8, no. 1, p. 53, Mar. 2021, doi: 10.1186/s40537-021-00444-8.

[14] V. Dissanayake, S. Seneviratne, R. Rana, E. Wen, T. Kaluarachchi, and S. Nanayakkara, “SigRep: Toward Robust Wearable Emotion Recognition With Contrastive Representation Learning,” *IEEE Access*, vol. 10, pp. 18105–18120, 2022, doi: 10.1109/ACCESS.2022.3149509.

[15] C. Szegedy *et al.*, “Going Deeper with Convolutions.” arXiv, Sep. 16, 2014. Accessed: Oct. 26, 2023. [Online]. Available: http://arxiv.org/abs/1409.4842

[16] V.-T. Ninh *et al.*, “An Improved Subject-Independent Stress Detection Model Applied to Consumer-grade Wearable Devices.” arXiv, Mar. 17, 2022. Accessed: Oct. 26, 2023. [Online]. Available: http://arxiv.org/abs/2203.09663

[17] R. Li and Z. Liu, “Stress detection using deep neural networks,” *BMC Med. Inform. Decis. Mak.*, vol. 20, no. S11, p. 285, Dec. 2020, doi: 10.1186/s12911-020-01299-4.

[18] P. Siirtola, “Continuous stress detection using the sensors of commercial smartwatch,” in *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, London United Kingdom: ACM, Sep. 2019, pp. 1198–1201. doi: 10.1145/3341162.3344831.

[19] M. Amin, K. Ullah, M. Asif, H. Shah, A. Mehmood, and M. A. Khan, “Real-World Driver Stress Recognition and Diagnosis Based on Multimodal Deep Learning and Fuzzy EDAS Approaches,” *Diagnostics*, vol. 13, no. 11, p. 1897, May 2023, doi: 10.3390/diagnostics13111897.

[20] P. Siirtola and J. Röning, “Comparison of Regression and Classification Models for User-Independent and Personal Stress Detection,” *Sensors*, vol. 20, no. 16, p. 4402, Aug. 2020, doi: 10.3390/s20164402.

[21] D. Lopez-Martinez, N. El-Haouij, and R. Picard, “Detection of Real-world Driving-induced Affective State Using Physiological Signals and Multi-view Multi-task Machine Learning.” arXiv, Jul. 19, 2019. Accessed: Oct. 28, 2023. [Online]. Available: http://arxiv.org/abs/1907.09929

***Extra text:*** *Heart rate (HR) is defined as the number of times the heart beats per minute, while heart rate variability (HRV) represents the time intervals between consecutive heartbeats.*