Deep Learning Analysis of Mobile Physiological, Environmental and Location Sensor Data for Emotion Detection

Eiman Kanjo[a], Eman M. G. Younis[b], Chee Siang Ang[c]

[a] *Computing and technology Department, Nottingham Trent University, Nottingham*
[b] *Faculty of Computers and Information, Minia University, Minia, Egypt*
[c] *School of Engineering and Digital Arts, University of Kent*

Abstract

The detection and monitoring of emotions are important in various applications, e.g. to enable naturalistic and personalised human-robot interaction. Emotion detection often require modelling of various data inputs from multiple modalities, including physiological signals (e.g.EEG and GSR), environmental data (e.g. audio and weather), videos (e.g. for capturing facial expressions and gestures) and more recently motion and location data. Many traditional machine learning algorithms have been utilised to capture the diversity of multimodal data at the sensors and features levels for human emotion classification. While the feature engineering processes often embedded in these algorithms are beneficial for emotion modelling, they inherit some critical limitations which may hinder the development of reliable and accurate models. In this work, we adopt a deep learning approach for emotion classification through an iterative process by adding and removing large number of sensor signals from different modalities. Our dataset was collected in a real-world study from smart-phones and wearable devices. It merges local interaction of three sensor modalities: on-body, environmental and location into global model that represents signal dynamics along with the temporal relationships of each modality. Our approach employs a series of learning algorithms including a hybrid approach

[*] Corresponding author
Email address: eiman.kanjo@ntu.ac.uk (Eiman Kanjo)

using Convolutional Neural Network and Long Short-term Memory Recurrent Neural Network (CNN-LSTM) on the raw sensor data, eliminating the needs for manual feature extraction and engineering. The results show that the adoption of deep-learning approaches is effective in human emotion classification when large number of sensors input is utilised (average accuracy 95% and F-Measure=%95) and the hybrid models outperform traditional fully connected deep neural network (average accuracy 73% and F-Measure=73%). Furthermore, the hybrid models outperform previously developed Ensemble algorithms that utilise feature engineering to train the model average accuracy 83% and F-Measure=82%)

Keywords: deep learning, emotion recognition, convoltutional neural network, long short-term memory mobile sensing

1. Introduction

The growing popularity of sensors and low power integrated circuits, together with the increasing use of wireless networks have led to the development of affordable, robust and efficient wearable devices which can capture and transmit data in real time for a long period of time. These data sources provide a unique opportunity for innovative ways in recognising human activities through human physiological sensing while also taking into account other natural environmental factors, such as weather, noise levels, etc. This could potentially contribute to better management of chronic diseases such as diabetes, asthma and cardiovascular diseases [1]. For example, extensive research has focused on automatic detection of physical exercises which are linked to a range of health related issues[2]. Due to these potential impacts, research work is on the rise with many algorithms being developed for a range of application areas in healthcare (e.g. symptoms monitoring, home-based rehabilitation) and beyond (e.g. security, logistics supports) [2, 3]. Some of these machine learning algorithms include multivariate regression, K-nearest Neighbour (KNN) classification combined with Dynamic Time Warping (DTW), etc. In addition, given the importance of mental health and its increasing impact on societies, researchers are now finding ways to accurately detect human emotion with the hope to develop intervention strategies for mental health and to provides rich contextual information which can be used to better understand mental health issues [4]. Furthermore, there have also been significant interests in emotion detection in human-computer interactions [5] due to its potential use, allowing us to design intelligent computer systems which are adaptable according to users emotional states, ensuring convergence and optimisation of human-computer interaction. Therefore, there have been numerous attempts to exploit machine learning techniques utilising sensor datasets for automatic emotion detection [6, 7, 8, 9]. To date, a significant amount of research in automatic emotion detection has been carried out primarily using visual, audio and movement data (e.g. facial expression, body postures, speeches) [6, 3, 8, 9, 10]. With the increased availability of low-cost wearable sensors (e.g. Fitbit, Microsoft writs bands), there is an emergence in research interest in using human physiological data (e.g. galvanic skin response (GSR), heart rate (HR), electroencephalography(EEG), etc.) for emotion detection. Apart from these, given the intimate links between emotion and environmental factors [6], studies are starting to look into using environmental sensors data and location patterns to infer human emotion [6]. Despite the possibility of sensing a wide range of information (from human physiology to environment), automatic human emotion classification remains very challenging due to the idiosyncrasy and variability of human emotional expressions [11]. The range of modalities of emotion expression could be very broad, with many of these modalities still being inaccessible to current sensor technology (e.g. blood chemistry). Many accessible physiological signals may be non-differentiable in emotion detection [11]. Furthermore, studies in automatic emotion detection rely on controlled samples in lab settings, where specific emotions are artificially triggered using audio-visual stimuli (e.g. presenting photos or videos to participants) or by asking participants to carry out carefully designed tasks to induce emotional states [12]. Although this type of controlled studies is valuable for certain applications (e.g. clinical diagnosis in healthcare), its use is rather limited to strictly controlled environments. For emotion detection technology to be useful in the everyday management of mental health and mobile human-computer interaction in the wild, we are interested in techniques which allow us to detect emotion on-the-go and in real-life settings. In this paper, we explore a deep learning approach for multivariate time series classification, combining environmental, physiological and location sensor data using smart phones and wristbands. Inspired by the deep feature learning in images and speech recognition [13, 14, 15], we explore a deep learning framework for multivariate time series classification for emotion recognition in the wild, where users are walking in a urban area. Deep learning relieves the burden of manually extracting hand-crafted features for machine learning models. Instead, it can learn a hierarchical feature representation from raw data automatically. We leverage this characteristic by building models using a range of deep learning methods to train raw sensor data. This eliminates the need for data pre-processing and feature space construction, and simplifies the overall machine learning process [16]. Due to its success in image and speech classification, deep learning has been increasingly used for non-image/speech data, including human activity recognition using time series data such as in the case of smart phone accelerometer data [17, 18, 1, 19]. There have also been recent attempts using deep learning for emotion detection, although most studies have only looked at lab based emotion data [20, 9]. Specifically, we propose a Multi-Channels Deep Convolutional Neural Network (MC-DCNN) model. We follow a hybrid approach based on Convolutional Neural Network and Long Short-term Memory Recurrent Neural Network (CNN-LSTM) inspired by previous state of the art [19, 21] which have been applied to human activity using accelerometer data. The majority of the studies employing deep learning on activity recognition are restricted to a handful of data channels as opposed to this study where we utilise 20 sensor channels from three different modalities to classify emotion against self-reported emotion labels. The main contribution of our work lies in:

1. The use of multimodal sensor feeds (physiological, environmental and lo-

cation data) for emotion detection using features automatically extracted with deep learning approach. Although deep learning has been used in human activity/emotion detection, few studies looked into multimodal datasets. Specifically, to the best of our knowledge, no other work has applied deep learning on the combination of physiological, environmental and location data for emotion recognition.

2. The collection of real-world data from participants walking in a transited city location wearing a wristband and smart phone, while reporting their emotion periodically using a smart phone. The data therefore better reflect the complexity of real life environments. Most previous studies in automatic emotion detection are carried out in controlled lab settings as opposed to "in the wild" (i.e. in participants' natural environments), therefore the results are restricted to narrow application domains.
3. Various experiments carried out to compare different architectures of deep neural networks, including hybrid models using hybrid multi-channel sensor data (beyond human activity recognition).
4. The analysis and fusion of human physiological, environmental and location features individually and combined to explore its significance in emotion classification.

2. Related work

In recent years, smart phones and many wearable devices such as smart watches and wristbands are equipped with a range of sensors which can continuously monitor human physiological signals (e.g. heart rate, motions/movements, location data) and in some cases the ambient environment data (e.g. noise, brightness, etc.). This led to the emergence of large datasets in a wide variety of research areas such as healthcare and smart city. This burst of on-Body and environmental data presents an unprecedented opportunity for healthcare research, but it requires the development of new tools and approaches to deal with large multidimensional datasets. In the past decades, researchers from various fields, particularly in ubiquitous and mobile computing have been exploring the possibilities harnessing these data to infer or predict human behaviour, with varying levels of success [17, 22, 23, 18, 1, 19, 24, 25]. Given the relative ease of collecting time series sensor datasets, researchers have investigated the relationship between these sensor data and human emotion. The majority employ traditional statistical analysis methods and machine learning techniques. Often, a number of hand-crafted features that summarise the raw sensor data are extracted from the less structured data. These features are then filtered empirically or using structured algorithms through a feature selection process [16]. Features with low level of correlation with its corresponding label are excluded (through dimensionality reduction). Moreover, features are often removed to avoid collinearity, when excessive correlation among explanatory variables (features) exist in the dataset. Given the list of selected features, computational models are built which help classify or predict human activity/emotion using machine learning models such as logistic regression based models [6], support vector machines (SVM), decision trees, artificial neural networks (ANN), etc. [26, 27]. Although hand-crafted features have yielded promising results, they are domain-specific, and often poorly generalise to other similar problem domains. Handcrafted-based approaches involve laborious human intervention for selecting the most discriminating features and decision thresholds from sensor data. Handcrafted features have a decisive impact on models [16] and often utilise statistical variables, e.g., mean, variance, kurtosis and entropy, as distinctive representation features. Moreover, traditional machine learning and feature engineering algorithms may not be efficient enough to extract the complex and non-linear patterns generally observed in multimodal time series datasets. In addition, traditional feature engineering could also result in a large output features set [28]. This is problematic because it is difficult to know without training which features are relevant to a given task, and which are noise. As a result, the ability to select features from a huge feature set is critical and will require additional dimensional reduction techniques to process these features. Furthermore, feature extraction and feature selection are computationally expensive.

The computational cost of feature selection may increase combinatorially as the number of features increases [28]. In general, search algorithms may not be able to converge to optimal feature sets for a given model [16]. Given the complexity of human emotion detection, it is important to have abstract representations of the sensor data which are invariant to local changes in the data. Learning such invariant features is a major challenge in pattern recognition (for example learning features which are invariant to the time of data collection). Traditional shallow methods, which contain only a small number of non-linear operations, do not have the capacity to accurately model such variation of time series data. Therefore, to overcome the difficulties in obtaining effective and robust features from time-series data, many researchers have turned their attention to deep learning approaches. One interesting property of deep learning techniques is that they can work on raw data and automate the feature extraction and selection. Noisy time series samples are fed into the network as input data, and during each transformation, a hidden representation of inputs from the prior layer is generated to form a higher hierarchical architecture of data representation (i.e. features). One can train the network by adjusting the mapping parameters, in order to obtain finer abstraction levels. Specifically, each layer in a deep learning model combines outputs from the previous layer and transforms them via a non-linear function to form a new feature set. This gives a deep learning model the ability to automatically learn features directly from the underlying sensor data, forming a hierarchy where basic features are detected in the first layers, and in the deeper layers the abstract features from previous layers are combined to form complex feature maps. Empirical studies showed that data representations obtained from stacking up non-linear feature extracting layers as in deep learning often yield better results, e.g., improved classification model accuracy [18], better generative models (to produce better quality samples) [18], and the invariant characteristics of data representations [18]. Deep learning techniques have already made significant impacts in computer vision [13, 29, 30], speech recognition [31, 32] and natural language processing [20, 33, 34] where it performs better than standard machine learning methods and the performance is comparable to human level. While some attempts at detecting human activity and emotion have been made using deep learning[17, 21, 20],[8, 9], it is still a new and growing area of research which requires further work. In recent years, deep learning has been increasingly used in the field of human activity recognition [17, 21]. While progress has been made, human activity recognition remains a challenging task. This is partly due to the broad range of human activities as well as the rich variation in how a given activity can be performed. Since deep learning is capable of high-level abstraction of data, it can be used to develop self-configurable frameworks for human activity as well as emotion recognition. For instance, in an attempt to improve performance accuracy of activity recognition using mobile phone triaxial accelerometer data, [17] utilised a hybrid approach of deep learning and hidden Markov models(HMM). This approach allows to model deep hierarchical representations of spatial data with restricted Boltzmann machines (RBM) and stochastic modelling of temporal sequences in the HMM models. The proposed approach was reported to have performed better than traditional methods of using shallow networks with handcrafted features. Other deep learning architectures, including Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) have been increasingly applied in activity recognition problems. The performance of CNN for some activity recognition tasks was explored by [35, 36, 37]. Building on CNNs successes in image recognition, [13] developed a method based on CNN and applied it in activity recognition problems in three different domains:assembling line activities, activities in kitchen and jogging/walking. CNN was utilised to automatically extract features from accelerometer data without any domain knowledge. It was shown that CNN can capture local dependencies and invariant features in the data. Experimental results showed that CNN outperformed traditional machine learning approaches. Using a CNN model, [17, 21] demonstrated that it can model complex multivariate sensory time series data (considering accelerometer and gyro data) in recognition common human activities, e.g. walking, sitting, laying, etc. Specifically, CNN outperformed SVM which has previously achieved the best performance in this dataset. [37] showed that CNN also outperformed other conventional machine learning methods (e.g. KNN and SVM) in two other activity recognition datasets: breakfast activity and gesture recognition. A CNN is used in [38] to extract features for gait pattern recognition so that labour intensive handcrafted feature extraction process is avoided. Furthermore, CNNs have been applied for detection of stereotypical movements in Autism [39], where they significantly improved upon the state- of-the-art. Recurrent neural network (RNN) relying on Long Short-Term Memory (LSTM) cells have gained popularity due to its ability to exploit the temporal dependencies in time series data. LSTM have recently achieved impressive performance in various time-dependent applications, such as machine translations, automatic video subtitling, and others [40]. A biometrics application of LSTM has been explored by [41] to identify individual humans based on their motion patterns captured from smartphones, i.e. accelerometer, gyroscope and magnetometer. This is a challenging task, as temporal motion signals are generally very noisy. Their work using LSTM demonstrated that human movement convey necessary information about the persons identity and it is possible to achieve relative good authentication results. Furthermore, the same LSTM algorithm can also be applied to other time series data on gesture detection in a human conversation. In [21] a hybrid approach was used based on CNN and LSTM to classify human activities using two public datasets (daily activities and assembly line activities). The fundamental idea is to use CNN to automatically extract spatial features from raw sensor signals, and LSTM to capture the temporal dynamics of the human movement. Their results showed that CNN-LSTM hybrid model outperformed other deep models without using LSTM to model time dependencies. Importantly, it was shown that the model can potentially be used in multimodal sensor data.

2.1. Convolutional Neural Networks (CNN)

Convolutional neural networks (CNN) are a widely used deep learning algorithm which performs especially well for images input data, although they are now increasingly applied in time series data including human physiological data and financial data [42, 21]. The inputs in a convolutional layer connect to the subregions of the layers instead of being fully-connected as in traditional neural networks models. These group of inputs in subregions share the same weights, therefore the inputs of a CNN produce spatially-correlated outputs, whereas in traditional neural networks (NN), each input has individual weight and hence produce independent outputs. In a neural network with only fully-connected layers, the number of weights can increase quickly as the dimension of the input increases. CNNs reduce the number of weights compared with NN with the reduced number of connections through weights sharing and downsampling. CNNs typically consist of of three types of layers: convolutional layers, pooling/downsampling layers and fully-connected layers.

- The convolutional layer is the main building block of a CNN which determines the output of connected inputs in within local subregions. This is done via a set of learnable filters (kernels) which are convolved across the the width and height of the input data, calculating the scalar product between the values of the filter and the input, hence producing a two dimensional activation map of that filter. Through this, CNNs are able to learn filters which activate when specific type of features at some spatial position of the input are detected.

- The pooling layer will perform downsampling along the spatial dimensionality of the given input, further reducing the number of weights within that activation.

- The fully-connected layers are standard deep neural networks and attempt to produce predictions from the activation, to be used for classification or regression.

Convolution is the key operation in CNN. By convolution of the input signal with a linear filter (or kernel), adding a bias term and then applying a non-linear function, a 2D matrix named feature map is obtained, representing local correlation across the input signal. Specifically, for a certain convolutional layer, the units in it are connected to a local subregion of units in the (l-1)th layer. Note that all the units in one feature map share the same weight vector (for kernel) and bias, hence, the total number of parameters is much less than traditional multilayer fully connected neural networks with the same number of hidden layers. This indicates that CNN has a sparse network connectivity [14], which results in considerably reduced computational complexity compared with the fully connected neural network. For a richer representation of the input, each convolutional layer can produce multiple feature maps. Though units in adjacent convolutional layers are locally connected, various salient patterns of the input signals at different levels can be obtained by stacking several convolutional layers to form a hierarchy of progressively more abstract features. For the jth feature map in the lth convolutional layer Cl,j, the unit at the mth row and the nth column is denoted as vm,nl,j and the value of vm,nl,j is defined by:

$$\begin{aligned} vm, nl, j = \sigma(bl, j + \sum k \sum pa = 0Pl, a1 \sum pb = 0Pl, & \\ b - 1wpa, pbl, j, kvm + pa, n + pbl1, k) & \\ \forall n = 1, 2, , Nl, m = 1, 2, , Ml & \end{aligned} \tag{1}$$

where Ml and Nl are height and width of feature map Cl,j. bl,j is the bias of this feature map, k indexes over the set of feature map in the (l–1)th layer, wpa,pbl,j,k is the value of convolutional kernel at position (pa,pb), Pl,a and Pl,b are the size of the convolutional kernel, and $\sigma()$ is the Rectified Linear Units (ReLU) nonlinear function. ReLU is defined by:

$$\sigma(x) = max(0, x) \tag{2}$$

The proposed convolution operation is performed without zero padding (unlike the conventional approaches of image processing). This means each dimension of feature map will be reduced after a convolution operation. Thus:

$$Ml = Ml - 1 - Pl, a + 1Nl = Nl - 1 - Pl, b + 1 \tag{3}$$

where l is the index of the layer that performs convolutional operation.

2.2. Recurrent Neural Networks (RNN)

In a traditional neural network (with only fully connected layers) we assume that all inputs are independent of each other. In CNN, we have seen that inputs can be grouped into subregions where features are spatially dependent on each other and share the same weights. For some classification/learning tasks, the inputs are temporally dependent. For instance, if we want to predict the next word in a sentence, it is important to know which words came before it. RNNs can perform the classification task for every element of a time sequence, with the output being depended on the previous computations. Another way to think about RNNs is that they have a memory which captures information about what has been calculated so far. In other words, RNNs take as their input not just the current input data they see, but also what they perceived one step back in time. The decision a RNN reached at time step $t-1$ affects the decision it will reach at time step t. Hence, RNNs have two sources of input, the present and the recent past. Here is what a typical RNN looks like: In theory RNNs can make use of information in arbitrarily long sequences, but in practice they have difficulties learning long-range dependencies due to the vanishing gradient problem [43]. The vanishing gradient problem is the result of RNN seeking to establish connections between the final output and inputs from many time steps before as a RNN passes through many stages of multiplication. To address this, we adopt Long Short-term Memory (LSTM) as the RNN memory unit. LSTMs help preserve the error that can be backpropagated through time and layers by using a gated cell which determines what information from the prior step should be forgotten and what information in current time step should be remembered into the next state, via gates that open and close (activate and deactivate). This allows a RNN to continue to learn over many time steps, thereby opening a channel to link causes and effects remotely. 1

The structure of a LSTM cell is illustrated in Figure and the mechanism of the gates is described as follows: The first step in a LSTM cell is to decide what information we will forget from the cell state. This decision is made by a Sigmoid layer called the forget gate layer. It looks at h_{t-1} and x_t, and outputs

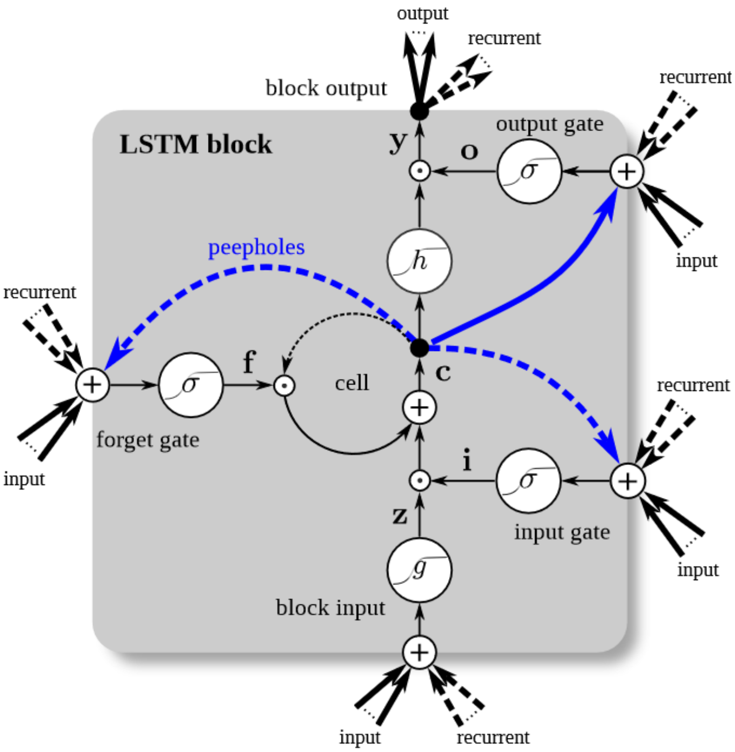

Figure 1: Long short term memory (LSTM) block [44].

a number between 0 and 1 for each number in the cell state C_{t-1}. 1 represents completely keep this while 0 represents completely remove this. The output ft of the gate is formalised as:

$$f_t = \sigma(Wf \cdot [h_{t-1}, x_t] + bf) \tag{4}$$

Then the cell decides which new information will be stored in the cell state. This has two parts. First, a sigmoid layer known as the input gate layer decides which values will be updated. Then, a tanh layer creates a vector of new candidate values, $\hat{C}_t$, which could be added to the state. These two will be combined to create an update to the state, as follow:

$$it = \sigma(Wi \cdot [h_{t-1}, x_t] + b_i) \tag{5}$$

$$\hat{C}t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{6}$$

Then, we update the old cell state, C_{t-1}, into the new cell state C_t as follow:

$$C_t = C_{t-1} * ft + i_t\hat{C}_t \tag{7}$$

To produce the output, a Sigmoid layer is first run, which decides which parts of the cell state will be output. Then, the cell state is fed through tanh (to push the values to be between -1 and 1) and multiplied by the output of the Sigmoid gate, as follow:

$$ot = \sigma(Wo \cdot [h_{t-1}, x_t] + b_o) \tag{8}$$

$$h_t = \tanh(C_t) * o_t \tag{9}$$

As we used Softmax as our last activation, our loss function is cross entropy loss:

$$Loss = -\sum_i log \frac{exp(Wx_i)}{\sum_j exp(Wx_j)} \tag{10}$$

Finally, Adam Optimizer can be used to have a better navigation through the loss function.

3. Methodology

In this section, we explain in details the dataset used for emotion in the wild classification and the architectures of deep learning models used for experimentation.

3.1. The EnvBodySens DataSet

We use the EnvBodySens dataset [6] to evaluate the models, which consists of 40 data files collected from 40 female participants (average age of 28) walking around the city centre in Nottingham, UK on specific routes. The dataset is composed of on-body data such as heart rate (HR), galvanic skin response (SGR), body temperature, motion data (accelerometer and gyro), environmental data such as noise levels, UV, air pressure and location data, GPS locations associated with time stamp and self report emotion levels (5-step Self-Assessment-Manikin (SAM) Scale for valence) logged by the EnvBodySens mobile application on Android phones (Nexus), connected wirelessly to Microsoft wrist Band 2 [45]. The participants were asked to spend no more than 45 minutes walking in the city center. Data was collected in similar weather conditions (average 20 ° degrees), at around 11am. During the data collection process, 550,432 sensor data frames were collected as well as 5,345 self-report responses. The statistical data analysis of the dataset is reported in [6]. Participants were asked to periodically report how they feel based on predefined emotion scale as they walked around the city centre. We adopted the 5-step SAM Scale for Valence taken from [46] to simplify the continuous labelling process. On average, 134 self-reports were entered per participants. We disabled the screen auto sleep mode on our mobile devices, so the screen was kept on during the data collection process. [6]. Data from six users were excluded due to logging problem. For example, one user was unable to collect data due to battery problem with the mobile phone, another user switched the application off accidentally. The correlation matrix in [6] shows a low level of correlation between the independent variables, suggesting that our model will not be affected by the multi-collinearity problem.

4. Model implementation

We use TensorFlow [47] to implement our models and Tensorboard for visualisation on Xeon E5-2640 v4 Processor (25Mb Cache, 2.4GHz, 8 core). In this paper, we first train a Multi-layer Perceptron (MLP) for emotion classification based on twenty raw sensor input from three modalities: i) on-body (i.e. physiological and motion/movement data), ii) environmental, iii) and location data. Initially, we train each modality individually and then we combine all sensor input modalities in a separate training process, see Figure 2 for the four different learning architectures. Then we evaluate the performance of each modality against the combined model. This is then followed by training deep learning models in order to test the efficacy of the deep learning approach for accurately classifying multimodal time series data.

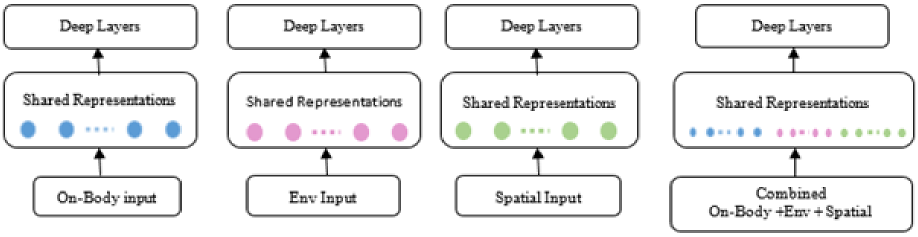

Figure 2: Learning Architectures, four models are trained, a) for On-Body, b) for Env , c) for Location separately d) and then fused using all the data input and feed it into the Deep Layers.

4.1. Data pre-processing

After the data collection the signals were pre-processed and cleaned. The first and the last 30 s were removed from the start of the data collection process for each user data, the reason for this step is that users needed a few seconds to fully engage in the movement and also few seconds to terminate the data collection process. A non-overlapping sliding window strategy has been adopted to segment the time series signal.12 shows the difference between the two segmentation methods.

4.2. MLP Models

Our implementation of ”Multi-Layer Perceptron” (MLP) network consists of two hidden layers. The first layer has 64 neuron, whereas the second hidden layer has 32 neurons. The input layer is 20*40 dimensions per iteration. The output layer has 5 neurons, each corresponds to the 5 emotional classes.

4.3. CNN Models

We start with the notations used in the CNN. A sliding window strategy is adopted to segment the time series signal into a (n, c, t) tensor, where $n =$ number of instances, $c =$ sensor channels, $t =$ time steps. After preliminary experiments with various deep learning topologies using multiple modalities combinations, we choose the CNN architecture as follows: Input of n batch x 20 channels x t window size, 2 convolutional layers (Conv1, Conv2), 2 maxpooling layers (Pooling1, Pooling2) and fully-connected layer as shown in 3. The first layer Conv1 has 32 filters (feature maps) while the second one Conv2 has 64 filters. This procedure may hinder partially the generality of the created models, as the average cross-validation accuracy is used to guide the feature selection search. However, the comparison between single, multiple modalities, and across fusion approaches is fair because all experiments follow the same procedure. The window size, r=40 (i.e. the height of sliding window) is chosen experimentally, by trying different sample rates from 10 to 100 as shown later in table 2. The convolution kernel is 2x2, stride is 1x1, i.e. strides= [1,1,1,1], Padding=1 (which does not shrink the matrix).

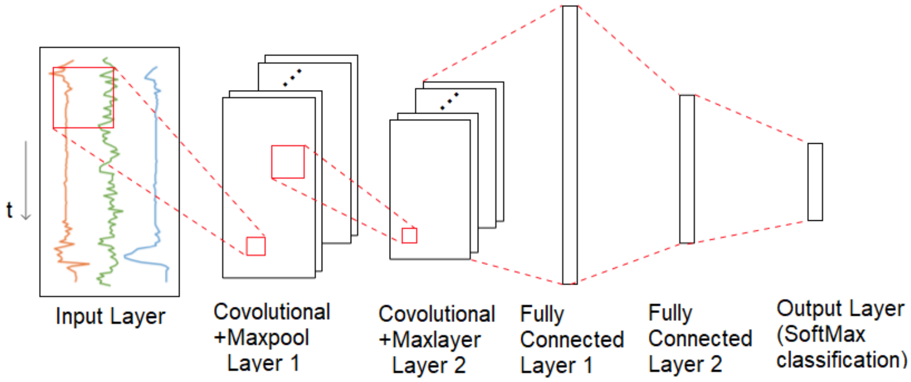
t
Input Layer
Covolutional +Maxpool Layer 1
Covolutional +Maxlayer Layer 2
Fully Connected Layer 1
Fully Connected Layer 2
Output Layer (SoftMax classification)

Figure 3: CNN architecture

4.4. CNN-LSTM Models

CNN-LSTM has a similar structure as CNN, with an added LSTM layer (see Figure3). In particular, the temporal dimension of the data is preserved during the convolution operation, and the resulting fully connected layer is fed into LSTM cell(see Figure 4). Each LSTM cell keeps track of an internal state that represents its memory. Over time the cells learn to output, overwrite, or reset their internal memory based on their current input and the history of past internal states. The MaxPool kernel is 2x2, stride is 1x2, i.e. strides= [1, 1, 2, 1], padding is 1. So that the temporal dimension is preserved and we only shrink

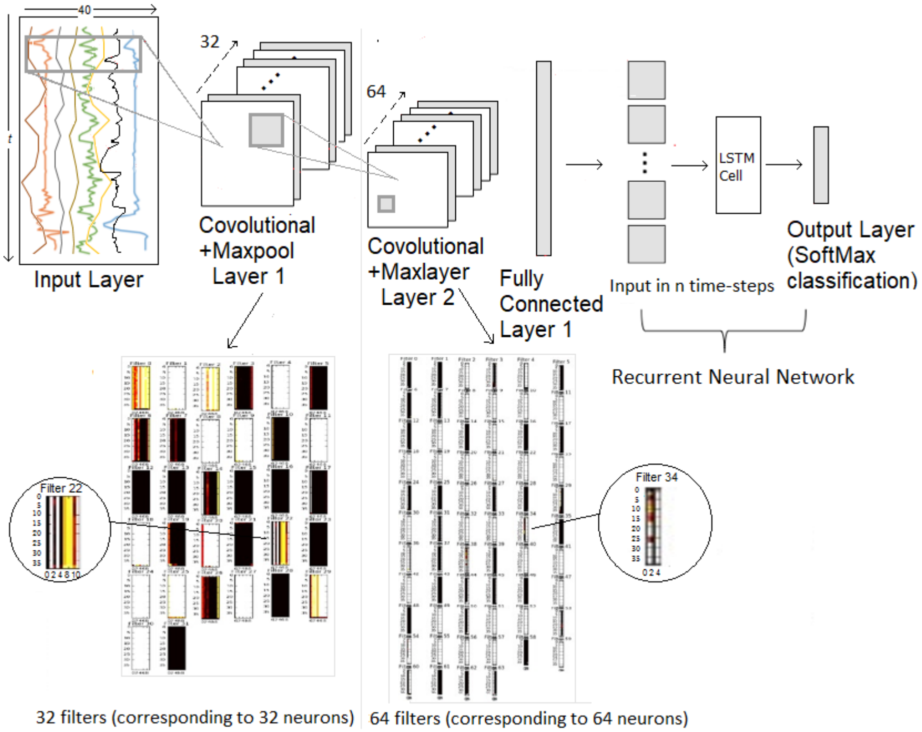

Figure 4: CNN-LSTM Model Architecture, we train 4 models separately, these are a) On-Body, b) Env , c) for Location d) and then we fused model using all the data input and feed it into the Deep Layers

the spatial dimension. Thus output filters are of 40x10 dimension after first MaxPool layer and 40x5 after the second MaxPool layer. Similarly, 32 filters are used for the first Conv1 layer and 64 used for the second Conv2 layer. The output of these filters are also shown in Figure 4. We train all models mentioned above on each subject dataset using fused data from all sensors modalities. We also train the models on three subsets of the data based on three modalities: on-Body, Env and Location. In total, we train 12 models on each user dataset (3x3 models on subsets and 3 models on fused data). Here, n=40 raw samples and c=20 the number of the attributes from sensor input. Similarly, c=2 for the location models, c=3 is for Env models (Noise, Air-Pressure and UV) and c =15 for the On-body models (the rest of the attributes).

Table 1: Average Performance metrics for all the models

	Average	Precision	Recall	F-measure	Accuracy	RMSE
MLP	All	0.734	0.728	0.729	72.9	0.95975
	Body	0.654	0.621	0.63	62.2	1.264
	Environment	0.424	0.428	0.424	42.6	1.54
	Location	0.59	0.605	0.58	60.2	1.22
CNN	All	0.818	0.79	0.787	78.6	0.788
	Body	0.734	0.712	0.709	70.8	1.01
	Environment	0.529	0.47	0.468	46.5	1.41
	Location	0.79	0.761	0.769	78.7	0.99
CNN-LSTM	All	0.927	0.95	0.949	94.7	0.291
	Body	0.881	0.878	0.874	87.3	0.6
	Environment	0.607	0.593	0.574	59.7	1.18
	Location	0.655	5.586	0.621	64	1.03

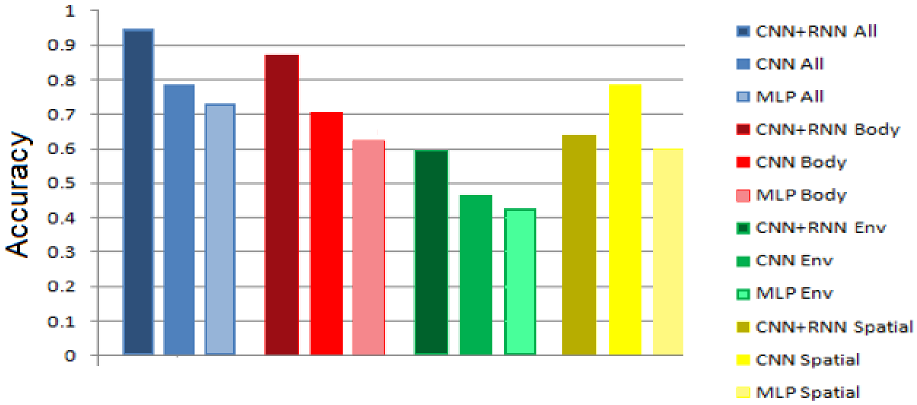

Figure 5: Comparison of average accuracy levels of all models

4.5. Results

All the experiments presented here are run for data files of each individual participant and then the average (and standard error) of the resulting models prediction accuracy and other performance metrics are reported. The performance of the trained model is evaluated by splitting each subject data using random sampling technique into training set of 70% data instances and test set of 30%. Evaluation results across all experiments are illustrated in Table 1, based on five standard performance evaluation metrics: Precision, Recall, F-Measure, Accuracy, Error rate and RMSE (root mean squared error). The accuracy levels of the results are also compared between single modalities (on-body, environmental and location modality) and combined modalities across all the three models. When MLP was trained only on on-Body data subset, it achieved an average accuracy of 0.62 (F-Measure: 0.63 ± 0.039). Location model achieved an average accuracy of 0.60 (F-Measure:0.580.032) while MLP did not not perform well on Environment data with an average accuracy lower than 0.50. MLP achieved an average accuracy of 0.72 (F-Measure:0.580.032) when performed on fused modalities data which is significantly higher than each single modality (p <0.01). Moreover, the results show that CNN outperforms MLP significantly by 6% (p <0.01) with an average accuracy 0.79. (F-Measure:0.79 ± 0.034). Both on-Body and Location models were improved with CNN. CNN-LSTM model achieved an average accuracy of 0.95 (F-Measure:0.95 ± 0.022) with significant 16% increase margin in performance, compared to CNN model (p <0.01). Furthermore, the accuracy level of the CNN-LSTM model increased considerably based on on-Body data at 0.87 (F-Measure:0.87 ± 0.024, (p <0.01)), although the model did not do as well with Location data. The results suggested that on-Body modality is the more robust data source for emotion classification ”in the wild”. The other two modalities, i.e. Environment and Location, are not as effective on their own but together they yield improved performance when fused with on-Body data by approximately 7% in accuracy. The high levels of accuracy achieved with the hybrid CNN-LSTM model reinforces the effectiveness of deep learning in multimodel time series sensor data for emotion recognition.

Due to limited space, we only visualise the accuracy levels of ten participants. Radar chart in Figure 5 shows the difference in accuracy levels of 10 users experiments which are selected randomly. With CNN-LSTM accuracy levels ranging between 0.89 to 0.996 (± 0.027). Similarly, Figure4.5, presents 3 radar charts of 10 users models (figure per model MLP,CNN and CNN-LSTM). Its clear from Figure4.5 that MLP models resulted in the highest variation between users, and models based on Environment data achieved the lowest accuracy levels. While in 4.5, we can see that all the combined modalities have achieved high levels of accuracy. Figure 4.5, illustrate the confusion matrices yielded by the three models based from one user data. There is a slight confusion between state 0 and 1 (negative emotions), which is improved when LSTM is added to the architecture. During the user study, we have made a great effort to ask users the meaning of each class and the difference between the very negative label ”0” and neutral ”3”. In addition, we have cropped the first few minutes of the data recording when users are stationary and using default rating (label) at 3. We believe our dataset is reasonably balanced with small variation from one user to another.

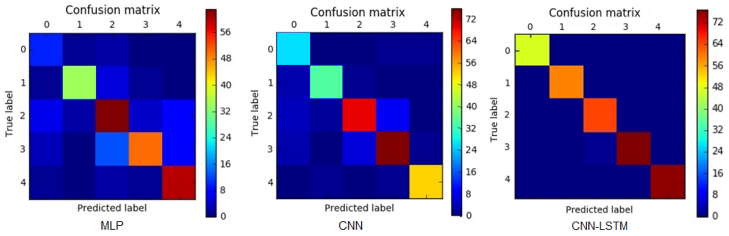

Figure 6: Confusion matrices of three models for one user data (fused data).

Modern deep learning techniques allow us to train a network in batches by interleaving multiple sequences together. Among others, batching allows to further exploit the power of matrix multiplication on the GPU and to avoid loading all data into memory at once. The batch size has implications for the robustness of the error that is propagated in the learning phase [29]. Figure 3 shows an example of 3 batches that encode 3 sequences of 5 samples each.

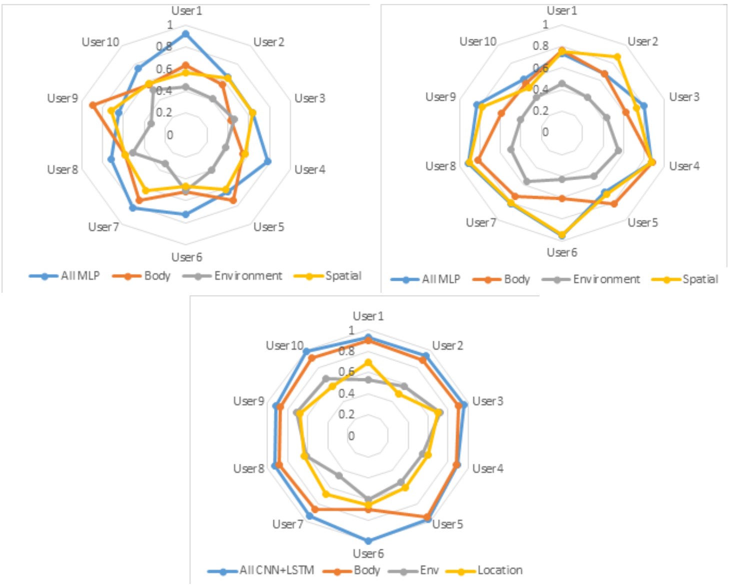

Figure 7: Radar charts showing the accuracy levels of three models(a) MLP, (b) CNN , and CNN=LSTM, based on ten users data in ad-hoc and fused modes.

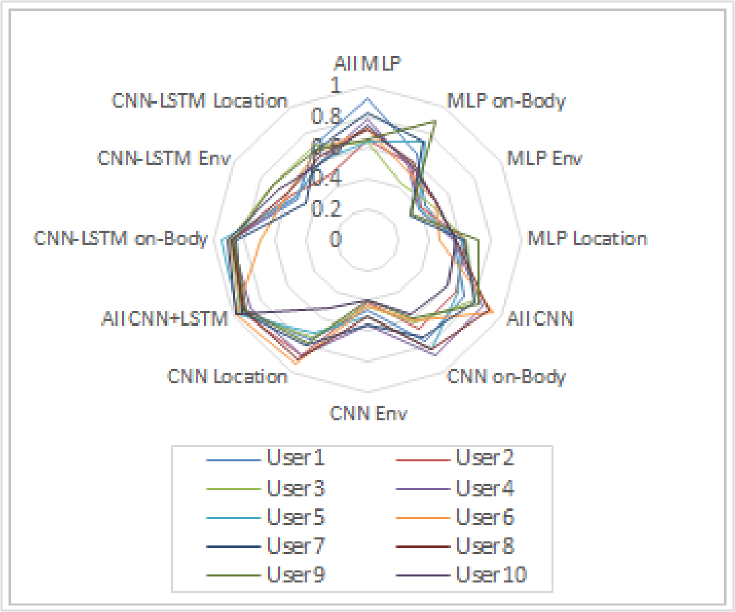

Figure 8: The accuracy levels of 10 users across all the models in ad-hoc and fused modes.

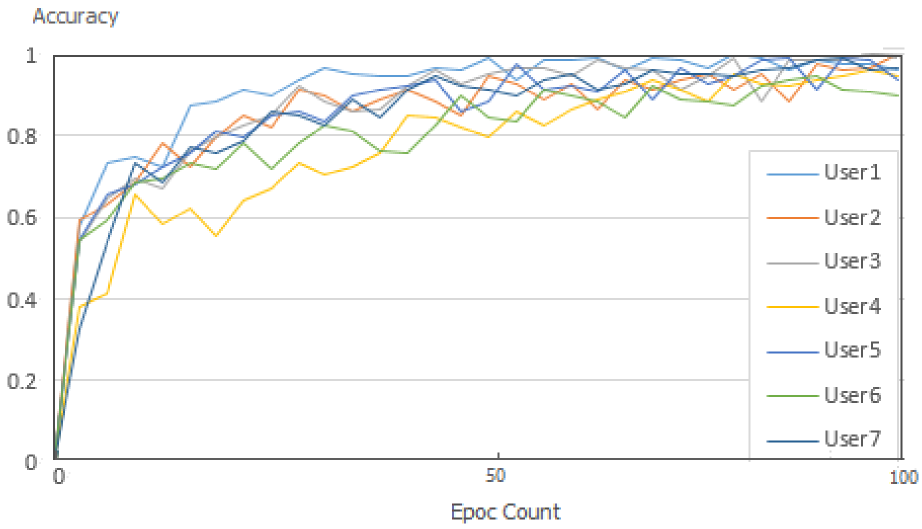

Figure 9: Cumulative distribution of recognition accuracy of 7 user.

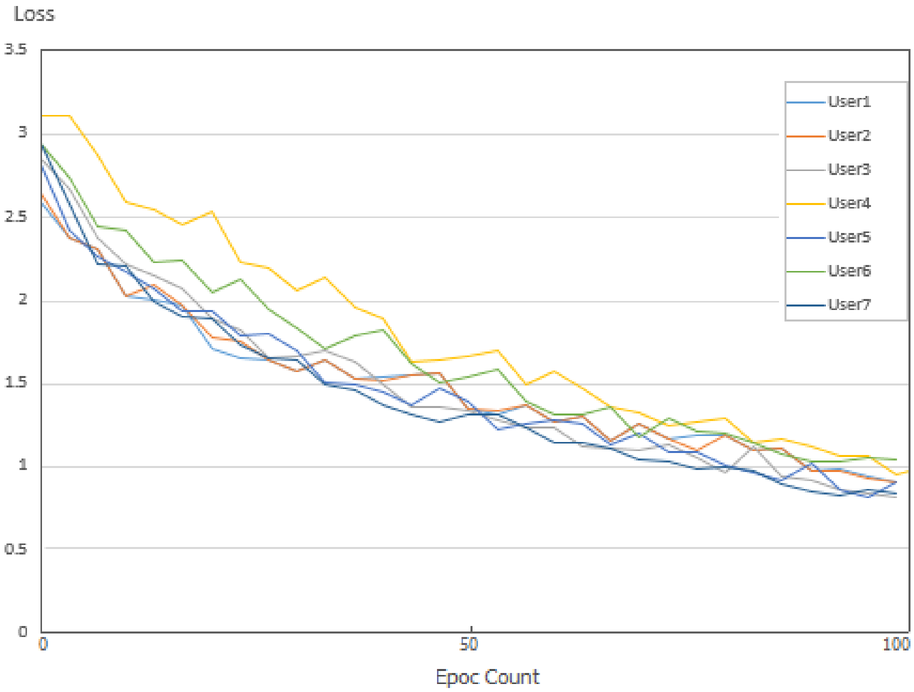

Figure 10: Cumulative distribution of recognition loss of 7 users.

5. Discussion

The objectives of our study were to (i) evaluate deep learning as a computational model for emotion recognition "in the wild" following state-of-the-art methodologies, and (ii) to assess the overall power of deep learning on multmodal sensor data including time series sensor input (Physiological, Environmental and Location data). This is one of the few studies looking into emotional recognition of participants in their natural environment using multiple sources of time series data. Our results have demonstrated that raw features can perform well when fused utilising deep learning models. In particular, CNN combined with LSTM has outperformed traditional MLP by more than 20% increase margin. Furthermore, applying deep learning on multimodel sensor data outperformed our earlier Ensemble algorithm by %6 margin [6] (see figure 11) which is based on staking various learners and refine the output by another meta learner layer.

Our results in general have suggested that deep learning methodologies are appropriate for modeling affective states and, more importantly, indicated that ad-hoc feature extraction may not be necessary for as deep learning models are able to identify high level of data abstraction automatically. Furthermore, in some affective states examined (e.g., relaxation models built on Electrodermal Activities (EDA); fun and excitement models built on Blood Volume Pulse (BVP); relaxation models built on fused EDA and BVP), deep learning without prior feature selection manages to reach or even outperform the performances of models built on ad-hoc extracted features which are boosted by automatic feature selection. These findings showcased the potential of deep learning for affective modeling based multiple sensors and multiple modalities input, as both manual feature extraction and automatic feature selection could be ultimately bypassed. Even though the results obtained are more than encouraging with respect to the applicability and efficacy of deep learning for affective modelling, there are a number of limitations and research directions that should be considered in future research. There are many parameters that can be tuned to obtain the optimal performance of the network. e.g. we have managed to test various

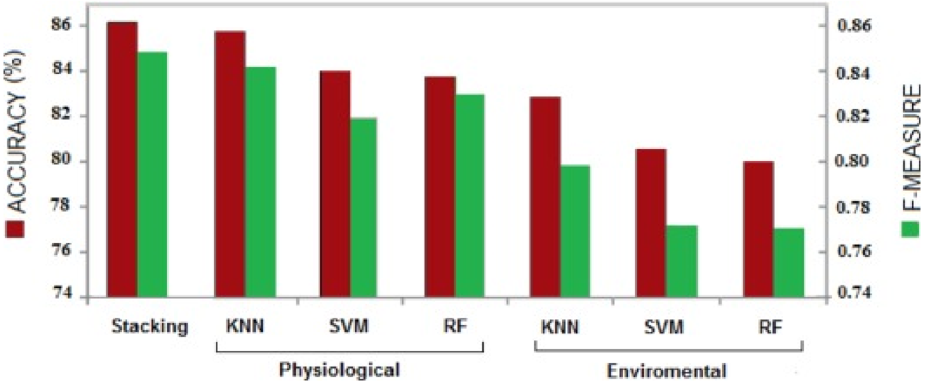

Figure 11: Accuracy and F-Measure levels of the base learners and the Stacking learner[6].

step sizes of the sliding window as shown in Figure 12. It demonstrates that by only analysing a small chunck of data (40 samples, i.e. 160ms), the deep learning model is able to classify emotions at high accuracy levels. The test has shown that the model performs at its best when the sliding window step size is set to 40. However, there are other parameters which can be tuned based on similar tests such as allowing window overlapping and the width of window overlap as shown in Figure 12. While the EnvBodySens dataset includes key components for emotion modelling and is representative of a typical affective modelling scenario, our approach needs to be tested on diverse datasets with larger number of participants and with more modalities and account to other factors such as pollution levels and crowd density, which may have significant impact on human emotions. Furthermore, we expect that the application of deep learning to model affect in large physiological datasets would show larger improvements with respect to statistical features and provide new insights on the relationship between physiology and affect. In addition, we have demonstrated that our algorithms can work on three very different modalities including physiological, enviromental and movement activities, we believe our models can also work on almost any other sensor data (beyond emotion detection and city sensing). Also we are in the process of deploying real-time mobile applications that can run these models on mobile and IoT platforms such as Intel Edison module [48].

Table 2: Average accuracy of CNN+LSTM models using different sliding window sizes. Bold numbers represent the best performing window size

Window Size	F-measure	Accuracy	RMSE
20	0.942	94	0.5
40	**0.949**	**94.7**	**0.291**
60	0.946	94.7	0.313
80	0.911	92.7	0.8
100	0.922	93.7	1.1
120	0.912	92.5	1.3

We have attempted to combine all participants data into one single dataset for emotion detection, however we found a high across-subject variation in the dataset which led to low model accuracy of less than 50%. This observation is in agreement with previous studies [49] which verify that emotion recognition is subject dependent which makes it difficult to obtain a generalised model across individuals.Others have successfully created a universal deep learning model for gesture data as gestures performed by different individuals are typically quite similar.For emotion however , there is higher levels of variations between individuals. Our results, confirm this, and verify that the emotion recognition is subject dependent as the accuracy varies from subject to subject and exhibits high variance of accuracy.

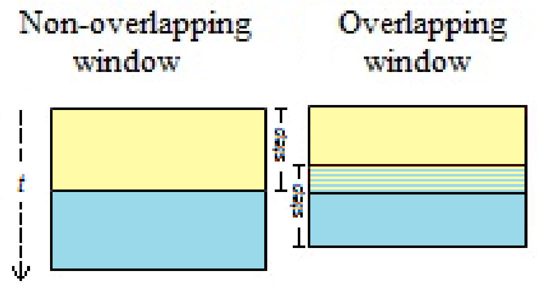

Figure 12: Illustration of sliding window steps and overlapping

6. Conclusion

Mobile phones along with other wearable devices produce large number of data as people are going about their daily activities. In this study, we presented a scenario of emotion detection ”in the wild” where people are moving from one place to another in an urban environment. Although this type of time series data can help us understand peoples emotion, traditional emotion recognition techniques requires features engineering process to be applied to data prior to modelling, which might be challenging especially if the dataset is multimodal and large. Deep learning offers an automated way for features extraction embedded in the process. This paper has demonstrated the advantages of employing a hybrid deep learning approach for raw multimodal data modelling based on smart device sensors input collected in city space. Our results have shown that using a hybrid deep learning approach (CNN-LSTM) on large number of raw sensor data increased the accuracy levels of emotion models by more than %20 compared to a traditional MLP model. Furthermore, fusing various sensor modalities including on-Body, Environment and Location data showed a significant increase in accuracy when compared to modelling single modality such as physiological sensors only. Also, we have shown that deep learning can be a promising approach for the study of human behaviour and emotion data. The promising results demonstrated in the study holds the potential for novel applications in emotion recognition and can open new opportunity in the study of mental health and well-being in real-life settings. In future work, we will further explore the possibility of utilising LSTM gates to reset and forget some of the states based on the emotion states and their history. Finally, we are planning to run a larger scale studies with other modalities and sensor feed such as(e.g. EEG data, air quality), and then build an emotion map using our model on mobile devices along with the sensors.

References

[1] G. Plasqui, K. R. Westerterp, Physical activity assessment with accelerometers: an evaluation against doubly labeled water, Obesity 15 (10) (2007) 2371–2379.

[2] O. D. Lara, M. A. Labrador, A survey on human activity recognition using wearable sensors., IEEE Communications Surveys and Tutorials 15 (3) (2013) 1192–1209.

[3] E. B. McClure, K. Pope, A. J. Hoberman, D. S. Pine, E. Leibenluft, Facial expression recognition in adolescents with mood and anxiety disorders, American Journal of Psychiatry 160 (6) (2003) 1172–1174.

[4] T. Pham, T. Tran, D. Phung, S. Venkatesh, Deepcare: A deep dynamic memory model for predictive medicine, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2016, pp. 30–41.

[5] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, J. G. Taylor, Emotion recognition in human-computer interaction, IEEE Signal processing magazine 18 (1) (2001) 32–80.

[6] E. Kanjo, E. M. Younis, N. Sherkat, Towards unravelling the relationship between on-body, environmental and emotion data using sensor information fusion approach, Information Fusion 40 (2018) 18–31.

[7] E. Kanjo, D. J. Kuss, C. S. Ang, Notimind: Utilizing responses to smart phone notifications as affective sensors, IEEE Accessdoi:10.1109/ACCESS.2017.2755661.

[8] S. Jerritta, M. Murugappan, R. Nagarajan, K. Wan, Physiological signals based human emotion recognition: a review, in: Signal Processing and its Applications (CSPA), 2011 IEEE 7th International Colloquium on, IEEE, 2011, pp. 410–415.

[9] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, S. Narayanan, Analysis of emotion recognition using facial expressions, speech and multimodal information, in: Proceedings of the 6th international conference on Multimodal interfaces, ACM, 2004, pp. 205–211.

[10] E. Kanjo, A. Chamberlain, Emotions in context: examining pervasive affective sensing systems, applications, and analyses, Personal and Ubiquitous Computing (2015) 1–16.

[11] R. W. Picard, Affective computing: challenges, International Journal of Human-Computer Studies 59 (1) (2003) 55–64.

[12] F. Agrafioti, D. Hatzinakos, A. K. Anderson, Ecg pattern analysis for emotion detection, IEEE Transactions on Affective Computing 3 (1) (2012) 102–115.

[13] G. E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural computation 18 (7) (2006) 1527–1554.

[14] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

[15] A. Graves, A.-r. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: Acoustics, speech and signal processing (icassp), 2013 ieee international conference on, IEEE, 2013, pp. 6645–6649.

[16] A. Supratak, C. Wu, H. Dong, K. Sun, Y. Guo, Survey on feature extraction and applications of biosignals, in: Machine Learning for Health Informatics, Springer, 2016, pp. 161–182.

[17] M. A. Alsheikh, A. Selim, D. Niyato, L. Doyle, S. Lin, H.-P. Tan, Deep activity recognition models with triaxial accelerometers., in: AAAI Workshop: Artificial Intelligence Applied to Assistive Technologies and Smart Environments, 2016.

[18] C. A. Ronao, S.-B. Cho, Human activity recognition with smartphone sensors using deep learning neural networks, Expert Systems with Applications 59 (2016) 235–244.

[19] N. Y. Hammerla, S. Halloran, T. Ploetz, Deep, convolutional, and recurrent models for human activity recognition using wearables, arXiv preprint arXiv:1604.08880.

[20] X. Zhou, J. Guo, R. Bie, Deep learning based affective model for speech emotion recognition, in: Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), 2016 Intl IEEE Conferences, IEEE, 2016, pp. 841–846.

[21] F. J. Ordóñez, D. Roggen, Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition, Sensors 16 (1) (2016) 115.

[22] N. Alajmi, E. Kanjo, N. El Mawass, A. Chamberlain, Shopmobia: An emotion-based shop rating system, in: Conference on Affective Computing and Intelligent Interaction, 2013, pp. 745–750. doi:10.1109/ACII.2013.138.

[23] L. Al-barrak, E. Kanjo, E. M. G. Younis, Neuroplace: Categorizing urban places according to mental states, 2017, PLOS ONE 12.

[24] W. Kieran, K. Eiman, Things of the internet (toi): Physicalization of notification, in: UbiComp '18 Proceedings of the 2018 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, ACM, 2018.

[25] W. Kieran, K. Eiman, Emoecho: A tangible interface to convey and communicate emotions, in: UbiComp '18 Proceedings of the 2018 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, ACM, 2018.

[26] M. Dumas, Emotional expression recognition using support vector machines, in: Proceedings of International Conference on Multimodal Interfaces, 2001.

[27] C.-C. Lee, E. Mower, C. Busso, S. Lee, S. Narayanan, Emotion recognition using a hierarchical binary decision tree approach, Speech Communication 53 (9) (2011) 1162–1171.

[28] J. Bins, B. A. Draper, Feature selection from huge feature sets, in: Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on, Vol. 2, IEEE, 2001, pp. 159–165.

[29] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, Greedy layer-wise training of deep networks, in: Advances in neural information processing systems, 2007, pp. 153–160.

[30] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[31] G. Dahl, A.-r. Mohamed, G. E. Hinton, et al., Phone recognition with the mean-covariance restricted boltzmann machine, in: Advances in neural information processing systems, 2010, pp. 469–477.

[32] G. E. Dahl, D. Yu, L. Deng, A. Acero, Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition, IEEE Transactions on audio, speech, and language processing 20 (1) (2012) 30–42.

[33] R. Socher, E. H. Huang, J. Pennin, C. D. Manning, A. Y. Ng, Dynamic pooling and unfolding recursive autoencoders for paraphrase detection, in: Advances in Neural Information Processing Systems, 2011, pp. 801–809.

[34] A. Bordes, X. Glorot, J. Weston, Y. Bengio, Joint learning of words and meaning representations for open-text semantic parsing, in: Artificial Intelligence and Statistics, 2012, pp. 127–135.

[35] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu, J. Zhang, Convolutional neural networks for human activity recognition using mobile sensors, in: Mobile Computing, Applications and Services (MobiCASE), 2014 6th International Conference on, IEEE, 2014, pp. 197–205.

[36] C. A. Ronao, S.-B. Cho, Deep convolutional neural networks for human activity recognition with smartphone sensors, in: International Conference on Neural Information Processing, Springer, 2015, pp. 46–53.

[37] J. Yang, M. N. Nguyen, P. P. San, X. Li, S. Krishnaswamy, Deep convolutional neural networks on multichannel time series for human activity recognition., in: IJCAI, 2015, pp. 3995–4001.

[38] M. Gadaleta, M. Rossi, Idnet: Smartphone-based gait recognition with convolutional neural networks, arXiv preprint arXiv:1606.03238.

[39] N. M. Rad, A. Bizzego, S. M. Kia, G. Jurman, P. Venuti, C. Furlanello, Convolutional neural network for stereotypical motor movement detection in autism, arXiv preprint arXiv:1511.01865.

[40] F. Yan, K. Mikolajczyk, Deep correlation for matching images and text, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3441–3450.

[41] N. Neverova, C. Wolf, G. Lacey, L. Fridman, D. Chandra, B. Barbello, G. Taylor, Learning human identity from motion patterns, IEEE Access 4 (2016) 1810–1820.

[42] M. Längkvist, L. Karlsson, A. Loutfi, A review of unsupervised feature learning and deep learning for time-series modeling, Pattern Recognition Letters 42 (2014) 11–24.

[43] R. Jozefowicz, W. Zaremba, I. Sutskever, An empirical exploration of recurrent network architectures, in: Proceedings of the 32nd International Conference on Machine Learning (ICML-15), 2015, pp. 2342–2350.

[44] K. Greff, R. K. Srivastava, J. Koutnk, B. R. Steunebrink, J. Schmidhuber, Lstm: A search space odyssey, IEEE Transactions on Neural Networks and Learning Systems 28 (10) (2017) 2222–2232.

[45] Microsoft Wrist Band kernel description, https://www.microsoft.com/microsoft-band/en-gb, accessed: 2017-09-04.

[46] E. Banzhaf, F. de la Barrera, A. Kindler, S. Reyes-Paecke, U. Schlink, J. Welz, S. Kabisch, A conceptual framework for integrated analysis of environmental quality and quality of life, Ecological Indicators 45 (2014) 664 – 668. doi:http://dx.doi.org/10.1016/j.ecolind.2014.06.002.
URL http://www.sciencedirect.com/science/article/pii/S1470160X14002532

[47] Abadi, Others, {TensorFlow}: Large-Scale Machine Learning on Heterogeneous Systems, arXiv preprint arXiv:1603.04467.
URL http://tensorflow.org/

[48] Intel Edison kernel description, https://software.intel.com/en-us/iot/hardware/edison, accessed: 2017-09-04.

[49] Y. Gao, H. J. Lee, R. M. Mehmood, Deep learninig of eeg signals for emotion recognition, in: Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on, IEEE, 2015, pp. 1–5.