2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI)

Deep-Accel: A Face Touch Prediction Framework to Reduce Obsessive-Compulsive Disorder

Samuel Fipps[*‡], Bai Chen[†§], Lisa Anthony[*§], Mamoun T. Mardini[†§], and Arunkumar Bagavathi[*‡]
Department of Computer Science[*], Department of Health Outcomes and Biomedical Informatics[†]
Oklahoma State University[‡], University of Florida[§]
sfipps@okstate.edu, chenbai@ufl.edu, lanthony@cise.ufl.edu, malmardini@ufl.edu, abagava@okstate.edu

***Abstract*—Obsessive-compulsive disorder (OCD) can come in multiple types among human beings. An important type of OCD is touching the face intentionally / unintentionally, particularly when in deep thinking. Examples include pulling hair, rubbing eyes, pinching pimples, vaping, and scratching the chin. Such a constant urge can easily transmit germs and may also become vulnerable for disabled and older adults. Although there are several work that uses machine learning algorithms to identify human activities, there is very limited work on detecting face touch with the raw accelerometer data. There are multiple challenges to using accelerometer data for machine learning algorithms like noise and data dimensionality. In this paper, we introduce a deep learning framework *Deep-Accel* to predict face touching activities compared to other human activities. We give an in-depth analysis of the proposed framework and show that it achieves up to 79.6% accuracy in classifying face touching activities and up to 73.2% in classifying all activities.**

***Index Terms*—face touching recognition, OCD mitigation, wearable sensors**

I. Introduction

Obsessive-Compulsive Disorder, or OCD, is a mental health condition that affects millions of people worldwide [1]. It is characterized by persistent, intrusive thoughts, known as obsessions, and repetitive behaviors or mental acts, known as compulsions [2]. These acts are often related to mental states like fear, anxiety, depression, happiness deep thinking, and anger. Body-focused repetitive behaviors (BFRBs), such as hair pulling, skin picking and nail-biting, are common habits [3] that involve face touching. Excessive face-touching not only poses psychological implications but also increases health risks, especially in times of infectious diseases [4]. Therefore, innovative interventions are needed to reduce compulsive face-touching behaviors among individuals with OCD.

Wearable technologies like smartwatches offer unique opportunities for health interventions in this digital age. Smartwatches are convenient and capable of real-time data collection and feedback, providing an invaluable platform for behavioral modification techniques [5]. Our study introduces a novel approach to mitigating obsessive face-touching behaviors using smartwatch data. We propose a face touching prediction framework that utilizes machine learning algorithms on movement data collected from smartwatches. By identifying high-risk instances before they occur, this framework can trigger timely interventions, aiding in the reduction of compulsive face-touching and alleviating symptoms of OCD. Our research aims to contribute to the growing body of knowledge on technological interventions for mental health disorders and present a scalable solution to detect face touch and treat individuals with OCD.

II. Related Work

A. Human Activity Recognition (HAR) with Machine Learning

Accurate recognition of human activities supports many downstream applications, such as elderly care, monitoring daily physical exercises, and monitoring workers' health in the working environment [6], [7]. Kwapisz et al. [8] used a multilayer perceptron to recognize 6 daily activities using accelerometer data collected from 29 participants by carrying a smartphone in their pockets. Zubair et al. [9] used random forest and AdaBoost to predict five states and transitions (sitting, sitting down, standing, standing up, and waking) from 4 participants who wore 4 tri-axial accelerometers at different body positions (waist, left thigh, right ankle, and right arm).

However, traditional machine learning methods applied in HAR rely heavily on extracting time- and frequency-domain features, limiting the ability of machine learning models to capture the temporal features from the data [10]. Deep learning approaches [11], such as CNNs and RNNs which automatically embody correlations between data that are close in the time series, were therefore applied to eliminate the constraints. Mardini et al. [10] combines three convolution layers and one LSTM layer to recognize the activity type (sedentary vs. locomotion vs. lifestyle) of 33 daily activities. Thama Alsarhan et al. [12] identified 17 activities (9 activities of daily living and 8 fall states) from data collected with the smartphone accelerometer from 30 participants using an RNN with bidirectional gated recurrent units.

Recently there have been research efforts to utilize machine learning algorithms to detect face touch using accelerometer information collected from smart watches [4]. Chen et al. used the dynamic time warping technique to detect face-touching actions using the accelerometer embedded in the smartwatches [13]. There were also attempts to apply deep learning methods in face-touching recognition, especially CNN and RNN, given the time-series output of the accelerometer [14].

B. OCD and Machine Learning

Recent advancements in wearable sensors and machine learning have shown promise in detecting OCD symptoms. Wahl et al. [15] conducted a study using data collected from

2835-5776/24/$31.00 ©2024 IEEE
DOI 10.1109/IRI62200.2024.00020

TABLE I: Table of actions in the dataset and whether they are considered face touching or not.

Activity	Touching Face?
Adjusting Eyeglass	YES
Repeated Face Touching	YES
Eating and Drinking	YES
Simulated Smoking	YES
Using Mobile Phone	NO
Leisure Walk	NO
Moving Items From One Location to Another	NO
Computer Tasks	NO
Lying Flat on the Back	NO
Writing	NO

accelerometers, gyroscopes, and magnetometers to distinguish between compulsive and routine hand washing in 21 participants. They compared the performance of four machine learning models (linear support vector machine (SVM), SVM with radial basis functions, random forest, and naïve bayes) using a window size of 10 seconds. Similarly, Lønfeldt et al. [16] investigated the feasibility of predicting OCD episodes in adolescents through multimodal biosensor data, including blood volume pulse, heart rate, electrodermal activity, and skin temperature. They extracted 66 features with a 5-minute window and evaluated the performance of four machine learning models (logistic regression, random forest, neural network, and mixed-effect random forest). Despite these promising findings, there remains a lack of research investigating the use of wearable sensors and analyzing deep learning learning techniques to detect a wider range of OCD symptoms.

III. Dataset

The dataset used in our experiments contains accelerometer data collected from 10 participants (50% females, 47.7 ± 24.7 years old) by wearing a Samsung Gear smartwatch [4], [13]. We used a smartwatch application to collect the tri-axial accelerometer data at a sampling rate of 30Hz. Participants performed a battery of ten activities at their own pace in a standardized laboratory settings while wearing the smartwatch, as shown in Table I. Participants performed each task repeatedly for 3 minutes before taking a break and moving on to the next task. All the activities were completed in one visit in a randomly generated order to avoid any temporal dependence on behavior. *This dataset is openly available on Github.* [1]

We processed and cleaned the data in such a way that each participant has only 4,800 samples per action, where each sample is an accelerometer reading represented in (x, y, z) format. We split each individual's accelerometer readings with an action window (epoch) measured in Hertz (*Hz*). An action window is the estimated amount of time for each action to take place. After trying various action windows, we picked the window length of 192 (*Hz*), or about 6.3 seconds. We picked 192 (*Hz*) instead of 180 (*Hz*) (which would be 6 seconds) because 192 (*Hz*) is the closest number that would divide 4800 samples into equal labels allowing us not to discard any data.

IV. Methodology

A. Formal Notations and Problem Statement

The raw accelerometer readings are represented as $X \in \mathbb{R}^{N\times3}$, where N is the total number of activities (data points) performed by k participants in the study. Each activity $n_i \in N$ can be associated with only one of the categories $C = \{c_1, c_2, \ldots c_p\}$ where p is the number of unique human activities like eating, walking, smoking, etc in the data. Note that our set of categories includes q human face touching activities, where $q \leq p$, which are challenging to predict with ML models.

Given the raw accelerometer data $\mathbb{R}^{N\times3}$, our aim is to build a learning function $f : \mathbb{R}^{N\times3} \longrightarrow \mathbb{R}^{N\times d}$ to extract contextual feature representations $\mathbb{R}^{N\times d}$ of size d for each activity $n_i \in N$ and $d >>> 3$. In other words, we learn contextual representations $\mathbb{R}^{N\times d}$ from 3-dimensional accelerometer data to predict the category of a given activity $n_i \in N$, including the face touch, using a classifier function $g : \mathbb{R}^{N\times d} \longrightarrow \mathbb{1}_C$. We develop these classifier and representation learning models to identify q face touch activities from the p human activities.

B. Proposed Deep-Accel framework

We propose a deep learning framework *Deep-Accel* to learn contextual feature representations of raw accelerometer readings obtained from smartwatches to detect human activities. Although the proposed method is similar to current works [8], [10], [12] to differentiate human activities in general, the main emphasis of the proposed method is to identify face touching activities to help in alleviating OCD symptoms. As depicted in Figure 1, the proposed *Deep-Accel* framework includes *5 blocks* to capture contextual representations from the raw accelerometer data $\mathbb{R}^{N\times3}$. We give a description of each block in our model below.

1) Block-1: We used block-1 of our model to learn the multi-dimensional features of the accelerometer data input. Block-1 consists of a fully connected (FC) or linear layer followed by a transformer layer with multi-head attention and another fully connected layer. We convert the 3-dimensional input to multi-dimensional features due to limitations in using attention heads in the transformer module. With multi-dimensional features, the attention heads get a wide range of features to map the relevance of each dimension in the input sequence to learn a rich representation of the input. First, we convert the 3-dimensional input representation to the size of 2048 with a fully connected layer. We used a transformer module [17] using multi-head attention, as given in Equation 1, and then linearly transform the rich representations using a fully connected layer.

$$A_h = softmax(\frac{Q_h.K_h^T}{d^k}).V_h \qquad (1)$$

Where A_h is attention distribution from head h, Q_h, K_h, and V_h are query, key, and value splits respectively for head h, and d^k is the dimension of input and output features. The attention distribution of each head is concatenated, and we

[1] https://github.com/ufdsat/FTCode

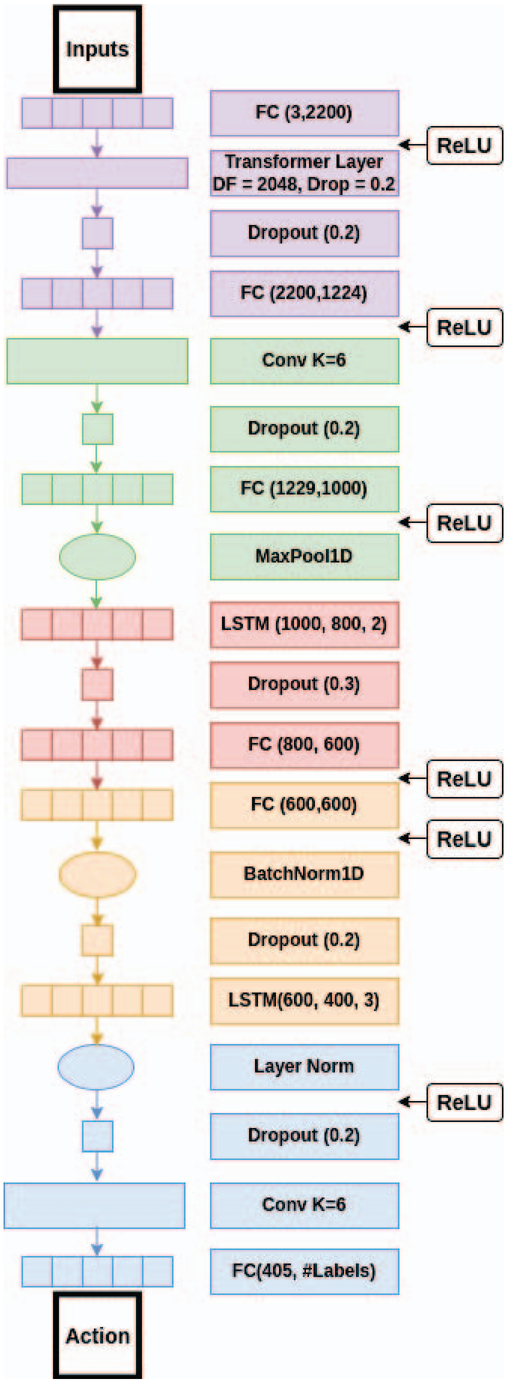

Fig. 1: Proposed *Deep-Accel* architecture with five blocks of neural networks that are designed to learn feature representations and predict human activities.

feed it to a fully connected layer. We used the ReLU activation function, given in Equation 2, for all fully connected layers.

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

2) *Block-2:* We used a Deconvolutional layer normally called transposed convolutional layer [18] in Block-2 as it is not always used to remove the effect of a convolutional layer. The transposed convolutional layer is considered an up-sampler of the input in that it generates feature representations with a higher number of dimensions compared to the input. We used a fully connected layer on the output of the transposed convolutional layer followed by a 1-dimensional power average max-pool as given in Equation 3.

$$f(x) = \sqrt[p]{\sum_{x \epsilon X} x^p} \tag{3}$$

3) *Block-3:* is designed to capture long-term dependencies, which is necessary for addressing the time series nature of the problem at hand. To achieve this, it employs a Long Short-Term Memory (LSTM) layer with a size of two. LSTM layers are particularly adept at handling time-dependent data and leverage three key variables to maintain and update information over time. The first of these variables is the cell state, which acts as the layer's long-term memory. The second variable is the hidden state or the previous state, and it works in conjunction with the third variable, the current state (representing the input data), to generate the layer's output. The LSTM's underlying formula guides this output computation. Afterward, the data passes through a dropout layer, followed by a fully connected layer, before proceeding to the next block of the model.

4) *Block-4:* Block-4 starts with a fully connected layer and passes through a batch normalization, as given in Equation 4 to compute normalization over 1 dimension and to reduce overfitting. γ is a scale parameter, β is a shift parameter with size equal to that of x, $E[b]$ is the mean of the given mini batch b, and $Var[b]$ is the variance of mini batch. This feeds into the next LSTM layer with a size of 3, allowing us to gather some more long-term dependencies.

$$BatchNorm(x) = \gamma * \frac{x - E[b]}{\sqrt{Var[b] + \epsilon}} + \beta \tag{4}$$

5) *Block-5:* is the last block in our model. We applied layer normalization to the input representation, a technique similar to Mini-batch normalization as given in Equation 4 except that it normalizes each input in a given layer independent of the batch. We increased the dimensions of the learned representations with a transposed convolutional layer and made predictions of actions with a fully connected layer. We learned the model parameters using a cross-entropy loss function $\mathcal{L}$ as given in Equation 5, where $p(x)$ is the probability of a class in the target and $q(x)$ is the probability of class in the prediction.

$$\mathcal{L} = -\sum_{\forall x} p(x) log(q(x)) \tag{5}$$

V. Experiments and Results

A. Experiment Strategies and Datasets

Although this work uses only one dataset, we created the following three types of datasets for our experiments.

1) Binary classification: We first evaluate the capacity of the proposed model to distinguish the accelerometer signals between face-touch and other actions in a binary classification setup. The dataset used in this experiment contains only two labels: *face-touch* (1) and *not face-touch* (0).

2) Face-touch actions classification: The second experiment for Deep-Accel is to test its capacity to distinguish different face-touch actions. For this experiment we considered only data from the four face-touch action categories.

3) All actions classification: This experiment measures the ultimate performance of the Deep-Accel. We considered all accelerometer signals available in our dataset with both face-touch and non-face-touch actions for this experiment.

B. Deep-Accel hyperparameters

1) Optimizer: We used Rectified Adam (*RAdam*) optimizer [19] to handle the convergence problem. We used the base parameters for RAdam, such as LR = 0.001, betas=(0.9, 0.999), and eps=1e-08. While training our model, we clip the gradients in every batch by taking the norm over all the gradients together and setting a norm max of 0.5 to prevent exploding gradients. The batch size used in our case was 1024 while using *Xavier* for weight initializations.

2) Block-1: : The first fully connected layer had an input of 3 and outputs of 2200. We set the transformer, the number of heads in the multi-head attention, and the number of dimensions in the feed-forward network of the transformer to 100 and 2048, respectively. The next layer is a fully connected layer with an input size of 2200 and outputs a 1224-dimensional representation.

3) Block-2: : The transposed convolutional layer in Block-2 was configured with 192 channels in and out with a kernel size of 6. This output is passed to a fully connected layer of input size 1229, and with an output size of 1000, that is passed through a ReLU function. This ReLU function leads to an LPPool1d(1,1).

4) Block-3: : This block starts with an LSTM layer that has an input size of 1000, a hidden size of 800 with 2 hidden layers, and a dropout equal to 0.3. We then apply a dropout of 0.3 again and pass to a fully connected layer with an input size of 800 and an output size of 600.

5) Block-4: : This block contains a fully connected layer with both input and output sizes set to 600. The output is passed through a 1-D batch normalization with a size of 192, which outputs to a ReLU function and another LSTM layer with an input size of 600 and 4 hidden layers of size 400.

6) Block-5: : Our final block passes through a layer norm of size 400 which is given to a transposed convolutional layer with 192 in and out channels with kernel size of 6. The output layer, which is a fully connected layer, has an input size of 405 and an output size depending on the number of classes.

TABLE II: Performance of Deep-Accel on three experiment strategies with traditional evaluation metrics

Model Name	**Accuracy %**	**F1 Score**	**Precision**	**Recall**
Binary classification	98.0%	0.983	0.976	0.979
Face-touch actions classification	79.6%	0.805	0.796	0.796
All actions classification	73.2%	0.735	0.732	0.727

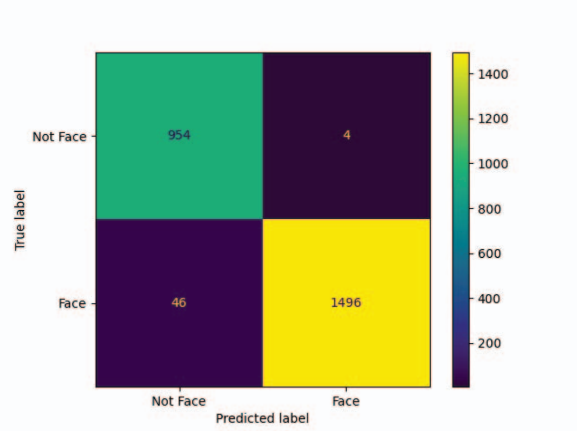

Fig. 2: Confusion matrix of binary classification experiment

C. Experimental results

Deep-Accel Performance on Classification tasks: Results of the proposed Deep-Accel are presented in Table II. The findings show a significant improvement in the model's performance compared to a baseline random prediction accuracy of 50%. The Deep-Accel model shows exceptional ability in distinguishing accelerometer signals related to face-touching and other activities, achieving a 98.0% accuracy as shown in row-1 of Table II.

However, it's worth noting a significant accuracy reduction of 18.4% in classifying face-touch actions specifically, as shown in row-2 of Table II. This indicates the complexity of accurately identifying face-touch actions and suggests the need for further advancements in machine learning methodologies to enhance the differentiation of such actions.

When considering the broader scope of all actions classification, the model achieves a 73.2% accuracy rate, as indicated in row-3 of Table II. This figure is notably aligned with the performance in face-touch action classification, despite the inclusion of a wider array of actions in the dataset.

Figure 2 shows the confusion matrix for the binary classification. It can be interpreted that the proposed model can correctly classify most of the accelerometer signals of both face-touch and other actions in the binary classification setting which supports the metrics given in Table II.

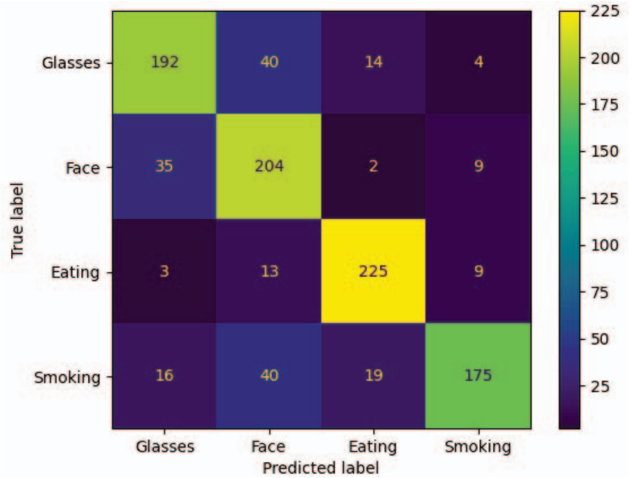

Fig. 3: Confusion matrix of face-actions classification experiment (each block is out of 250)

Figure 3 details the performance of Deep-Accel in classifying different face-touch actions. We can note that the proposed model mainly misclassifies ’adjusting glass’ actions to ’repeated face touch’ actions and vice versa. Similarly, ’smoking’ action is misclassified as ’eating’ and ’face touching’ which are almost the same set of actions in terms of accelerometer data. The proposed model predicts the majority of accelerometer data instances with an overall performance of 79.6%.

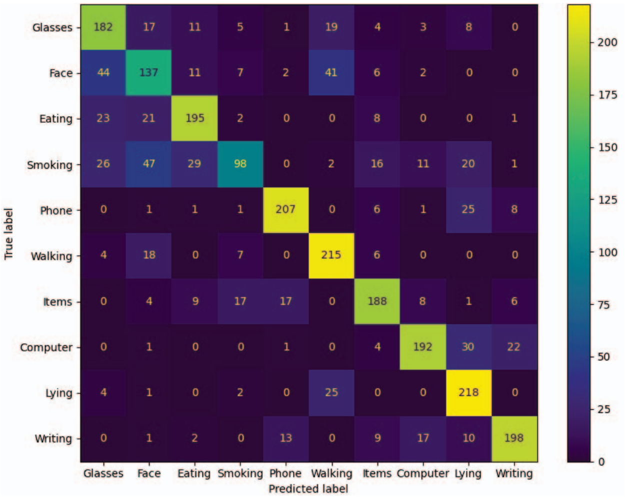

Fig. 4: Confusion matrix of all actions classification experiment (each block is out of 250)

TABLE III: Validation of the proposed Deep-Accel model by removing blocks to predict all activities

Block removed	Accuracy %	F1 Score	Precision	recall
No blocks	73.2%	0.735	0.732	0.727
Block1	64.2%	0.633	0.642	0.634
Block2	68.8%	0.689	0.688	0.680
Block3	66.4%	0.661	0.664	0.658
Block4	71.0%	0.713	0.71	0.701
Block5	70.56%	0.713	0.705	0.698

TABLE IV: Validation of the proposed Deep-Accel model by removing blocks to predict face touch

Block removed	Accuracy %	F1 Score	Precision	recall
No blocks	79.6%	0.805	0.796	0.796
Block1	65.5%	0.660	0.655	0.649
Block2	70.5%	0.711	0.705	0.703
Block3	71.0%	0.711	0.710	0.705
Block4	71.1%	0.709	0.711	0.706
Block5	73.7%	0.749	0.737	0.733

Figure 4 presents the confusion matrix of the proposed model for the all actions classification experiment. It is evident that misclassifications of the proposed *Deep-Accel* model in face touching actions affect the overall prediction performance. We can also note that some face-touch actions like ’*smoking*’ and ’*repeated face touch*’ have been mostly misclassified as one of the other actions. These patterns agree with the findings of our face touch action classification, shown in Figure 3. We find that accelerometer readings of instances that belong to some relaxed actions like ’*using computer*,’ ’*lying*,’ and ’*writing*’ also look similar and difficult to extract meaningful representations.

Ablation Study:

In addition to the general performance measures, we validated the impact of different modules and dataset samples on the proposed Deep-Accel model in this section. First, we removed blocks of the proposed model and showed the evidence of the importance of these blocks, and then validated the model performance for each individual present in the dataset to imitate the real-world scenario. Tables III and IV list the performance of Deep-Accel for actions classification and face touching actions classification experiments, respectively. Model performances given in these tables signify the importance of five blocks used in the Deep-Accel framework. Notably, we find that in both Tables III and IV the prediction accuracy of these experiments drop when removing one block from Deep-Accel compared to the prediction accuracy when none of these blocks are removed. Out of the five blocks, we note that the *block 1* is crucial for Deep-Accel in all predictions as the prediction accuracy gets 9% reduced and 14% reduced for all-actions classification and face-touch classification experiments, respectively. This, in turn, signifies the role of transformers [17] in accelerometer data prediction problems.

Table V lists the prediction accuracy of Deep-Accel for all three experiments for each individual present in the dataset. We run these experiments by considering all ten participants for model training and hiding two random actions for each person from the training data for model testing. We note that Deep-Accel for each individual performs at least as equally

TABLE V: Performance in Accuracy (%) of Deep-Accel at individual level on all dataset

Person Number	Binary Classification	Face-touch only	All actions
Person Number 1	99.6%	85.0%	78.4%
Person Number 2	99.2%	78.0%	77.2%
Person Number 3	99.6%	95.0%	88.8%
Person Number 4	100%	84.0%	82.8%
Person Number 5	100%	92.0%	87.6%
Person Number 6	87.2%	89.0%	53.2%
Person Number 7	100%	63.0%	77.6%
Person Number 8	67.2%	71.0%	34.0%
Person Number 9	92.8%	62.0%	76.4%
Person Number 10	95.6%	77.0%	76.0%

TABLE VI: Accuracy (%) of Deep-Accel at individual level for all actions excluding two people

Person Number	Accuracy Percentage
Person Number 1	79.6 %
Person Number 2	77.6%
Person Number 3	92.8%
Person Number 4	86.4%
Person Number 5	86.8%
Person Number 6	NA%
Person Number 7	74.8%
Person Number 8	NA%
Person Number 9	75.6%
Person Number 10	78%

as the general model for the overall dataset, except for two people - 6 and 8. We experimented with many parameters for the proposed Deep-Accel model and the performance for these two people is always significantly lower in all three experiments. Interestingly, we note from Table VI that when we exclude these two individuals completely from the model training and utilize data of only eight participants, the overall performance of Deep-Accel can increase a maximum of 4% for each individual.

VI. Conclusion and Discussion

Our study utilized deep learning to analyze accelerometer data from smartwatches for detecting face-touching activities, including those associated with OCD behaviors. The excellent performance of the models developed in this study indicates that wearable technology has the potential to detect OCD behaviors, which could be useful in developing intervention mechanisms and providing a tool for patients and clinicians. Our strategy can be refined in two ways. (i) Presently, we apply a uniform window length across all participants and labels, which has served as a starting point. However, tailoring this window length to align more closely with the timing of each action could further refine our label accuracy, providing a more nuanced understanding of the data. (ii) We also recognize the potential benefits of diversifying participant movement patterns in our study. While our current dataset benefits from a robust sample size from each participant, enhancing our participant base to include a broader spectrum of movement patterns could provide deeper insights.

References

[1] A. M. Ruscio, D. J. Stein, W. T. Chiu, and R. C. Kessler, "The epidemiology of obsessive-compulsive disorder in the national comorbidity survey replication," *Molecular psychiatry*, vol. 15, no. 1, pp. 53–63, 2010.

[2] J. N. Fenske and K. Petersen, "Obsessive-compulsive disorder: diagnosis and management," *American family physician*, vol. 92, no. 10, pp. 896–903, 2015.

[3] D. C. Houghton, J. R. Alexander, C. C. Bauer, and D. W. Woods, "Body-focused repetitive behaviors: More prevalent than once thought?" *Psychiatry research*, vol. 270, pp. 389–393, 2018.

[4] C. Bai, Y.-P. Chen, A. Wolach, L. Anthony, and M. T. Mardini, "Using smartwatches to detect face touching," *Sensors*, vol. 21, no. 19, p. 6528, 2021.

[5] D. J. Cook, M. Strickland, and M. Schmitter-Edgecombe, "Detecting smartwatch-based behavior change in response to a multi-domain brain health intervention," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 3, pp. 1–18, 2022.

[6] C. Bai, A. A. Wanigatunga, S. Saldana, R. Casanova, T. M. Manini, and M. T. Mardini, "Are machine learning models on wrist accelerometry robust against differences in physical performance among older adults?" *Sensors*, vol. 22, no. 8, 2022.

[7] A. D. Antar, M. Ahmed, and M. A. R. Ahad, "Challenges in sensor-based human activity recognition and a comparative analysis of benchmark datasets: A review," *International Joint Conference of ICIEV, and icIVPR*, pp. 134–139, 2019.

[8] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *SIGKDD Explor. Newsl.*, vol. 12, no. 2, p. 74–82, 2011.

[9] M. Zubair, K. Song, and C. Yoon, "Human activity recognition using wearable accelerometer sensors," in *2016 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*. IEEE, 2016, pp. 1–5.

[10] M. T. Mardini, S. Nerella, A. A. Wanigatunga, S. Saldana, R. Casanova, and T. M. Manini, "Deep chores: estimating hallmark measures of physical activity using deep learning," in *AMIA Annual Symposium Proceedings*, vol. 2020. American Medical Informatics Association, 2020, p. 803.

[11] S. Wan, L. Qi, X. Xu, C. Tong, and Z. Gu, "Deep learning models for real-time human activity recognition with smartphones," *Mobile Networks and Applications*, vol. 25, pp. 743–755, 4 2020.

[12] T. Alsarhan, L. Alawneh, M. Al-Zinati, and M. Al-Ayyoub, "Bidirectional gated recurrent units for human activity recognition using accelerometer data," in *2019 IEEE SENSORS*, 2019, pp. 1–4.

[13] Y.-P. Chen, C. Bai, A. Wolach, M. Mardini, and L. Anthony, "Detecting face touching with dynamic time warping on smartwatches: A preliminary study," in *Companion Publication of the 2021 International Conference on Multimodal Interaction*, ser. ICMI '21 Companion. ACM, 2021, p. 19–24.

[14] A. M. Michelin, G. Korres, S. Ba'ara, H. Assadi, H. Alsuradi, R. R. Sayegh, A. Argyros, and M. Eid, "Faceguard: A wearable system to avoid face touching," *Frontiers in Robotics and AI*, vol. 8, pp. 1–11, 2021.

[15] K. Wahl, P. M. Scholl, S. Wirth, M. Miché, J. Häni, P. Schülin, and R. Lieb, "On the automatic detection of enacted compulsive hand washing using commercially available wearable devices," *Computers in Biology and Medicine*, vol. 143, p. 105280, 2022.

[16] N. N. Lønfeldt, K. V. Olesen, S. Das, A.-R. C. Mora-Jensen, A. K. Pagsberg, and L. K. H. Clemmensen, "Predicting obsessive-compulsive disorder episodes in adolescents using a wearable biosensor—a wrist angel feasibility study," *Frontiers in Psychiatry*, vol. 14, p. 1231024, 2023.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] J. Qu, C. Su, Z. Zhang, and A. Razi, "Dilated convolution and feature fusion ssd network for small object detection in remote sensing images," *IEEE Access*, vol. 8, pp. 82 832–82 843, 2020.

[19] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," *arXiv preprint arXiv:1908.03265*, 2019.