Detection of Nocturnal Scratching Movements in Patients with Atopic Dermatitis Using Accelerometers and Recurrent Neural Networks

Arnaud Moreau, Peter Anderer, Marco Ross, Andreas Cerny, Timothy H. Almazan, and Barry Peterson

***Abstract*—Atopic dermatitis is a chronic inflammatory skin condition affecting both children and adults and is associated with pruritus. A method for objectively quantifying nocturnal scratching events could aid in the development of therapies for atopic dermatitis and other pruritic disorders. High-resolution wrist actigraphy (three-dimensional accelerometer sensors sampled at $\geq$20 Hz) is a noninvasive method to record movement. This paper presents an algorithm to detect nocturnal scratching events based on actigraphy data. The twofold process consists of segmenting the data into "no motion," "single handed motion," and "both handed motion" followed by discriminating motion segments into scratching and other motion using a bidirectional recurrent neural network classifier. The performance was compared against manually scored infrared video data collected from 24 subjects (6 healthy controls and 18 atopic dermatitis patients) demonstrating an F_1 score of 0.68 and a rank correlation of 0.945. The algorithm clearly outperformed a published reference method based on wrist actigraphy (F_1 score of 0.09 and a rank correlation of 0.466). The results suggest that scratching movements can be discriminated from other nocturnal movements accurately.**

***Index Terms*—Accelerometers, actigraphy, atopic dermatitis, long short-term memory, pruritus, recurrent neural networks, scratch.**

I. Introduction

ATOPIC dermatitis, along with other pruritic conditions, manifests itself in the sensation of itching resulting in scratching behavior. A practical method for quantifying the scratching behavior would greatly improve the ability to detect treatment efficacy in smaller clinical trials and thereby advance the drug development process.

The gold standard scratching assessment method is by scoring video recordings of patients [1], but this method is not practical for clinical trials or routine monitoring because of several limitations: its cost due to laborious scoring, the invasion of the patient's privacy and the possibility of visual obstructions [2]. Consequently, the current clinical standard for assessing scratching behavior in clinical trials is the use of patient reported outcomes (e.g. questionnaires) [3], [4] but these measures are imprecise and often do not agree well with objective measures [5].

Investigators have used a variety of objective measures to assess scratching including measurement of bed movements [6], finger flexation [7], sound detection [8], actigraphic assessment of sleep quality and quantity [9], [10], actigraphic measures of motion [2], [11]–[13], and attempts to actigraphically identify scratching events [14]–[16], but none of these methods have demonstrated sufficient accuracy and practicality in real-world situations to be accepted for routine use in clinical trials of therapies for pruritus. Therefore, the ability to quantify scratching in clinical trials remains an unmet medical need that hampers the development of effective therapies [17].

This paper describes the use of high resolution 3D actigraphy and recurrent neural networks to develop an assessment of nocturnal scratching that was developed and tested in a clinical setting against the gold standard measure with video recording.

Recurrent Neural Networks (RNNs) - specifically with long short-term memory architecture - have been used successfully to solve numerous tasks involving spatio-temporal data modeling (the interested reader may refer to [18] or [19] for a review). This type of model was frequently not only used to classify hand-crafted feature sequences, but also to learn feature representation directly from the raw data. Thus, the (often very tedious) process of finding suitably discriminating features that are extracted from the raw data in a precursory step can be replaced by the model directly learning features from the raw data automatically. Also in this work, the RNNs were applied directly to the resulting scratching event candidate raw data after the data were pre-segmented to determine whether motion was occurring at all and the motion was occurring in only one wrist or in both wrists simultaneously. We have employed Bi-directional RNNs with LSTM architecture described in [20], [21]. RNNs have been applied previously to accelerometer data in the context of gesture recognition in [22] (combining data from both an accelerometer and a gyrometer) or human activity recognition in [23], [24] (combining convolutional and recurrent units).

Manuscript received May 17, 2016; revised December 6, 2016, March 31, 2017, and May 16, 2017; accepted May 24, 2017. Date of publication June 7, 2017; date of current version June 29, 2018. This work was supported by Philips Respironics, Murrysville, PA USA. *(Corresponding author: Arnaud Moreau.)*

A. Moreau, P. Anderer, M. Ross and A. Cerny are with Philips, Sleep and Respiratory Care, Vienna 1120, Austria (e-mail: arnaud.moreau@philips.com; peter.anderer@philips.com; marco.ross@philips.com; andreas.cerny@philips.com).

T. H. Almazan is with Science37, Culver City, CA 90094 USA (e-mail: tim@science37.com).

B. Peterson is with Philips, Sleep and Respiratory Care, Bend, OR 97701 USA (e-mail: barry.peterson@philips.com).

Digital Object Identifier 10.1109/JBHI.2017.2710798

We propose a novel algorithm integrating data segmentation and RNNs that was evaluated on clinical data from 24 subjects (6 healthy controls, 18 atopic dermatitis patients) recorded in a sleep laboratory with IR video and compared against manual video scoring. Additionally, the algorithm proposed in [15] was run against the same data to provide baseline results for direct comparison.

This paper is organized in the following way: Section II addresses the data set and the algorithm that detects scratching episodes. In Section III the method is evaluated using a 6-fold cross-validation and the performance is compared to the algorithm outlined in [15]. Finally, Section IV discusses the obtained results.

II. Methods

This section describes the data collection and analysis and the scratching detection algorithm.

A. Data Sets

For training and validation of the algorithm a data set consisting of 24 recordings was collected. Participants wore 2 accelerometer devices (GeneActiv, Activinsights Ltd.) on each wrist for 2-5 nights, one of which they spent in a sleep laboratory while simultaneously being recorded on IR video. In this study, only the data collected during the night spent in the sleep laboratory was used. The raw 3-D accelerometer data from both wrists in units of g sampled at 100 Hz were read into Matlab (The MathWorks Inc., Natick, MA) as described in [25], synchronized with one another and down-sampled to 20 Hz.

The IR video was independently scored visually by an expert to identify individual scratching events that occurred throughout the night. A total of 24 participants completed the trial, 6 healthy controls (HC), numbered H001-H006 and 18 with atopic dermatitis (AD) numbered A001-A018, who fulfilled the Hanifin/Rajka diagnostic criteria [26] for atopic dermatitis. 9 were males and 15 were females. The average age of participants was 35.9 ± 15.1 and the average body mass index (BMI) was 25.5 ± 5.1. Of the 18 atopic dermatitis participants 7 were Asian, 5 were Caucasian, 5 were Hispanic or Latino and one identified as other. Of the 6 healthy volunteers 3 were Asian, one was Caucasian, one was African American and one identified as other. The study was approved by Schulman Institutional Review Board, number 201500381.

Scratching scoring was evaluated by IR video on the sleep lab night during the period in which the subject was lying in bed and at least attempted to sleep (e.g. excluding periods where subjects were reading or watching TV). The scoring was provided in Excel (Microsoft, Redmond, WA) files indicating start time, duration, dominant hand (left, right or both) and intensity (mild, moderate or severe). The scoring was synchronized to the actigraphy raw data using clapping movements executed by the subject on camera at the beginning and the end of the recording. Times where the subject was off-camera (e.g. bathroom breaks) and where the video scorer marked a scratching event as unclear were excluded from the analysis.

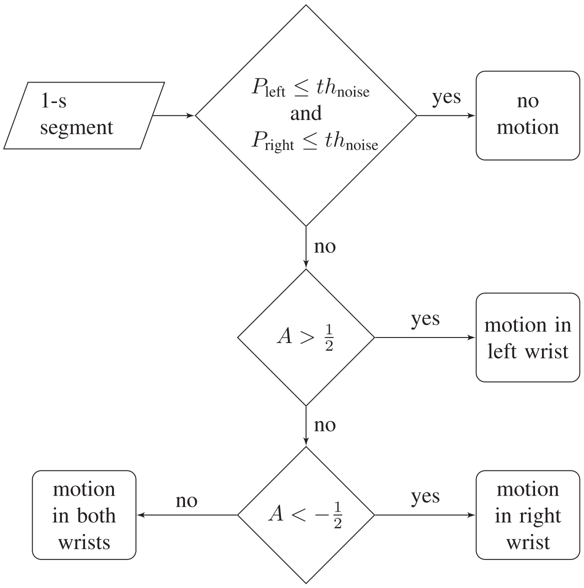

Fig. 1. Flow chart outlining the segmentation process, which was executed for each 1-s segment of data. P denotes the power calculated per wrist as defined in (1), A is the asymmetry index defined in (2) and th_{noise} denotes the background noise power level.

B. Segmentation

The purpose of the process described hereinafter was to identify scratching candidates (i.e. signal segments that contain motion caused by a single hand or both hands). Subsequently, these segments were classified as non-scratching motion or scratching motion (see Subsection II-D).

The acquired accelerometer raw data consisted of 3 dimensions denoted as $a_x(t)$, $a_y(t)$ and $a_z(t)$. It contained the acceleration caused by body movement along with the acceleration caused by gravity, which was assumed to be present in the low frequency components of the individual axis signals. Therefore, the gravity components $gr_x(t)$, $gr_y(t)$ and $gr_z(t)$ were extracted by applying a median filter of length 2 s to each dimension separately and then subtracting the resulting signal from the original $b_x(t) = a_x(t) - gr_x(t)$, $b_y(t) = a_y(t) - gr_y(t)$, $b_z(t) = a_z(t) - gr_z(t)$ to obtain an estimate of the body movement alone.

The next step was to segment the data into the following categories: "no motion", "single handed motion in the left wrist", "single handed motion in the right wrist" and "motion in both wrists" according to the flow chart depicted in Fig. 1. For this purpose, a power measure (mean vector magnitude) was derived for each consecutive 1-s window for both left and right wrist signals using the following equation:

$$P = \frac{1}{f_s} \sum_{t=1}^{f_s} \sqrt{b_x(t)^2 + b_y(t)^2 + b_z(t)^2}, \qquad (1)$$

where f_s denotes the sampling frequency. We empirically determined a minimum power threshold th_{noise} that needs to be exceeded in order for a segment to be considered as motion. Each segment (duration >3 s), where $P_{\text{left}} \leq th_{\text{noise}}$ and

$P_{\text{right}} \leq th_{\text{noise}}$ was thus excluded as no-motion (no scratching candidate). The next step was to separate asymmetric motion from motion in both wrists. For this purpose an asymmetry index for each consecutive 1 s window was calculated according to the following equation:

$$A = \frac{P_{\text{left}} - P_{\text{right}}}{P_{\text{left}} + P_{\text{right}}} \tag{2}$$

Segments where $|A| > \frac{1}{2}$ were assigned to either left wrist motion or right wrist motion depending on the sign. Gaps with duration <3 s in between left or right segments were closed. "Both wrist" candidates were defined as segments that belong neither to the left wrist nor right wrist motion category.

In order to increase the precision of finding meaningful "movement" borders, changes in hand orientation were estimated by calculating the derivative of the previously extracted gravity signals $gr_x(t)$, $gr_y(t)$, $gr_z(t)$ and then calculating the vector magnitude $C(t)$.

$$C(t) = \sqrt{\Delta gr_x(t)^2 + \Delta gr_y(t)^2 + \Delta gr_z(t)^2} \tag{3}$$

Δ thereby denotes a smoothed differentiation operator (implemented by discrete convolution). The underlying assumption is that the gravity vector $gr(t) = (gr_x(t), gr_y(t), gr_z(t))$ corresponds to the rotated gravitational field vector and thus indicating yaw, pitch and roll of the wrist. The threshold th_{orient} was empirically derived to describe a minimum change in hand orientation required to constitute a new "movement". Any local maximum $C_{\text{left}}(t) + C_{\text{right}}(t) > th_{\text{orient}}$ was considered to be a boundary of a scratching candidate segment. The new boundaries were applied additionally to the previous segmentation (left, right and both wrists).

C. Recurrent Neural Networks

RNNs differ from feed-forward neural networks by redirecting outputs back to inputs. This enables the model to consider context from the past (as well as the future when using bi-directional RNNs), which makes it especially suitable for modeling spatio-temporal data. Traditional RNNs however, suffer from the problem of exponential gradient decay (aka vanishing gradient problem) which limits the amount of context that can be considered severely [27]. Sepp Hochreiter *et al.* [28] proposed the so-called long-short-term-memory (LSTM) architecture as a solution to overcome this problem. Multiple additions to the LSTM architecture have been proposed since then, such as forget gates and peepholes [29].

Fig. 2 depicts a single LSTM unit. It consists of a memory cell controlled by gates, which enables the cell to learn when to recall the cell's content (output gate), when to update it (input gate) or when to overwrite it (forget gate). All the gate values are dependent on the same input data x_t and the previous outputs h_{t-1}, each associated with a different set of weights. The direct connections between the cell and its gates are known as "peepholes" and enable the gates to consider the cell's previous content c_{t-1}.

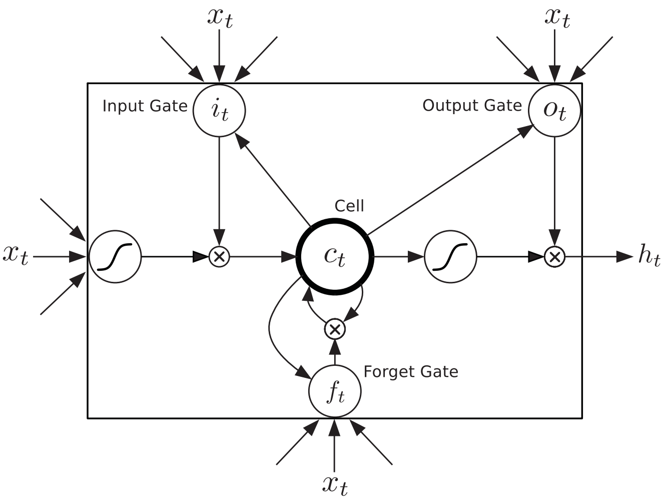

Fig. 2. The LSTM unit architecture (reproduced from [21]). Input data are propagated to 3 gates (input, output, forget) additionally to the input (on the left hand side). The cell corresponds to the memory that is updated on every time step (controlled by the gates). The output is shown on the right hand side.

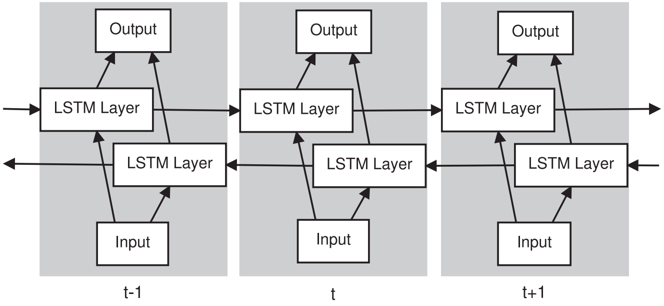

Fig. 3. Bidirectional RNN architecture unfolded in three illustrative time steps (adopted from [20]). The two LSTM layers in this architecture process the data in opposite directions.

D. Classification

We trained two bi-directional RNN (BRNN) models, one for asymmetric single handed movements, considering only the data from the active wrist, and another for both wrist movements. Any candidate scratching segment with duration $> th_{\text{dur}}$ was presented to the trained BRNN models for classification. th_{dur} thereby denotes the minimum duration of a segment to be considered a scratching candidate; the value of this threshold was empirically derived. Candidate segments with duration >3.5 s were further split into subsegments of approximate duration 3 s. The task of the models was to differentiate between non-scratching motion and scratching motion as no-motion segments have previously been excluded by the segmentation process (see Subsection II-B). Note that all empirically determined thresholds th_{noise}, th_{orient} and th_{dur} are set globally and thus do not vary by subject. The architecture of both models is depicted in Fig. 3. Each LSTM layer contains 3 (single wrist) or 4 (both wrists) LSTM units. The single wrist model input is $x_t = \big(b_x(t), b_y(t), b_z(t)\big)$ and the both wrists model input is $x_t = \big(b_x^{\text{left}}(t), b_y^{\text{left}}(t), b_z^{\text{left}}(t), b_x^{\text{right}}(t), b_y^{\text{right}}(t), b_z^{\text{right}}(t)\big)$, where the data has been normalized as determined during training (see Subsection II-F). The model output for each sequence consisted of a probability ($0 \leq p \leq 1$) that the motion sequence was caused by scratching. We applied a threshold of 0.5 to the output probability in order to obtain a final decision. The LSTM architecture included forget, input and output gates as well as peephole connections (for details see [21]). The total number of trainable weights was 193 (single wrist model) and 385 (both wrists model).

E. Cross-Validation

In order to use the entire 24 subject data set for both training and validation and still obtain an objective estimate of its performance on previously unseen data we performed a 6-fold cross-validation in the following way: The pre-segmentation procedure described in Subsection II-B was applied to all labelled recordings. Segments that were determined to be single handed scratching candidates and overlapped with a labelled scratching event were collected as scratching sequences. Analogously, segments that were determined to be single handed scratching candidates and did not overlap with a labelled scratching event were collected as no-scratching sequences. The both handed scratching candidates were handled in the same manner.

The collection of scratching and no-scratching sequences was further split into a test, training and validation data set. First, a test set was extracted, which contained 20% of all sequences chosen randomly from every subject, for the purpose of comparing models (independent of cross-validation folds). The remaining data were split randomly in 6 folds (each containing 3 AD and 1 HC subject). The performance was measured on 1 fold and the algorithm was trained on the other 5. This data set was divided by subject in order not to over-estimate the performance based on potential intra-subject scratching characteristic similarity.

To address the class imbalance problem [30], where the target classes are unequally distributed in the training, validation and test data sets, each set was re-sampled to achieve a distribution of 50% scratching sequences for single-hand sets and 33% scratching sequences for both-hands sets in the following way: for each subject, randomly chosen sequences from the majority class were removed proportionally to the target rate. If after this procedure too few sequences were left such that the target rate was too low, an adequate number of randomly chosen sequences that had been removed before were put back into the set.

F. Training

Training is the process of determining a set of weights for the BRNN model such that the error is minimized over the training data set. During the training, a given scratching candidate sequence $\mathbf{x}$ is presented to the input layer of the BRNN step-by-step and the activations are updated layer by layer (forward pass). Obtaining the net's output $y(\mathbf{x})$ for all training sequences $(\mathbf{x}, t) \in T$ (as described in Subsection II-E), where $t = 1$ or $t = 0$ depending on the target class (scratching or non-scratching), the cross-entropy error function [20] was computed by

$$E = -\sum_{(\mathbf{x},t)\in T} t \ln y(\mathbf{x}) + (1-t)\ln(1-y(\mathbf{x})). \tag{4}$$

The error was propagated backwards through the net (backward pass) by computing the partial derivatives of (4) using the standard back-propagation through time (BPTT) algorithm [31] and subsequently updating the network's weights using the stochastic gradient descent algorithm with momentum $= 0.9$ and learning rate $= 2 \times 10^{-4}$ [32]. Training was performed using the software RNNLib by Alex Graves [33]. The raw data have been normalized to have standard deviation 1 as suggested in [20, section 3.3.3] (note that the data had a mean ≈ 0 because of the gravity component removal). To prevent over-fitting the following two strategies were applied: Adaptive weight noise using the minimum description length error function [34] was used as a regularization mechanism during training. Also, the training was terminated when no improvements in the validation set error were found for the last 50 epochs (iterations through the training set). Typically, both the classification error over the training and validation set go down in the beginning, but as the training progresses, the error on the training set further reduces, whereas the error on the validation set stabilizes [20], which is when the model is over-fitting. The weights were initialized randomly (uniformly distributed in the range $[-1, 1]$, therefore this training procedure was run 32 times for both the single-hand and the both-hands network and the networks with the lowest error rate on the validation set were selected to be the final models.

G. Evaluation

Video and actigraphy scratching events were compared on a 1-s epoch basis, because video scratching was also scored on this time scale. The segmentation parameters were set to $th_{\text{noise}} = 0.02$ g (differentiating no movement from movement), $th_{\text{orient}} = 0.25$ g (refining segment boundaries) and $th_{\text{dur}} = 1.75$ s (determining the minimum segment length). Scratching events shorter than 2 s scored by video (in total about 1.6% of 1-s epochs) were removed. Ebata *et al.* [1] only used scratching bouts lasting more than 5 s to study the characteristics of nocturnal scratching in patients with atopic dermatitis. The clinical relevance of shorter bouts is an open research question.

In order to measure the performance of the algorithm to score scratching events based on high-resolution actigraphy data, its results were compared to the gold-standard measurement (IR video scored by experts). Three objectives were pursued when developing the algorithm (in order of increasing rigor):

1) The total duration of scratching in a recording obtained by the algorithm shall be as close as possible to the total duration of scratching obtained by the video scoring.
2) The separability between groups (atopic dermatitis and healthy controls) shall be maximized.
3) The scratching events obtained by the algorithm shall be placed as closely as possible to the scratching events obtained by the video scoring (temporal distribution).

The algorithm's performance was compared to the method detecting nocturnal scratching from actigraphy data published

TABLE I
PERFORMANCE MEASURES PER RECORDING RESULTING FROM 6-FOLD CROSS-VALIDATION AND, FOR THE PURPOSE OF COMPARISON, THE EQUIVALENT MEASURES FOR THE PETERSEN *et al.* ALGORITHM IN BRACKETS

													Scratching duration (s)			Total (s)
Rec	TP (s)		FP (s)		FN (s)		Sensitivity		Precision		F_1 score		Video	Actigraphy		duration
A001	1086	(37)	133	(79)	105	(1154)	0.91	(0.03)	0.89	(0.32)	0.90	(0.06)	1191	1219	(116)	24825
A002	445	(34)	59	(70)	245	(656)	0.64	(0.05)	0.88	(0.33)	0.75	(0.09)	690	504	(104)	25257
A003	278	(15)	73	(147)	70	(333)	0.80	(0.04)	0.79	(0.09)	0.80	(0.06)	348	351	(162)	23077
A004	124	(16)	182	(314)	104	(212)	0.54	(0.07)	0.41	(0.05)	0.46	(0.06)	228	306	(330)	23155
A005	735	(18)	383	(144)	250	(967)	0.75	(0.02)	0.66	(0.11)	0.70	(0.03)	985	1118	(162)	25294
A006	449	(45)	178	(259)	160	(564)	0.74	(0.07)	0.72	(0.15)	0.73	(0.10)	609	627	(304)	24935
A007	45	(9)	53	(171)	42	(78)	0.52	(0.10)	0.46	(0.05)	0.49	(0.07)	87	98	(180)	23533
A008	14	(1)	62	(141)	14	(27)	0.50	(0.04)	0.18	(0.01)	0.27	(0.01)	28	76	(142)	23311
A009	456	(27)	232	(237)	96	(525)	0.83	(0.05)	0.66	(0.10)	0.74	(0.07)	552	688	(264)	26677
A010	1318	(82)	618	(587)	681	(1917)	0.66	(0.04)	0.68	(0.12)	0.67	(0.06)	1999	1936	(669)	25469
A011	551	(25)	145	(109)	170	(696)	0.76	(0.03)	0.79	(0.19)	0.78	(0.06)	721	696	(134)	27454
A012	598	(90)	251	(192)	402	(910)	0.60	(0.09)	0.70	(0.32)	0.65	(0.14)	1000	849	(282)	25443
A013	3934	(557)	826	(422)	2968	(6345)	0.57	(0.08)	0.83	(0.57)	0.67	(0.14)	6902	4760	(979)	25198
A014	262	(8)	290	(109)	71	(325)	0.79	(0.02)	0.47	(0.07)	0.59	(0.04)	333	552	(117)	25509
A015	15	(0)	67	(59)	5	(20)	0.75	(0.00)	0.18	(0.00)	0.29	(0.00)	20	82	(59)	25159
A016	70	(3)	71	(115)	85	(152)	0.45	(0.02)	0.50	(0.03)	0.47	(0.02)	155	141	(118)	26296
A017	27	(4)	80	(70)	26	(49)	0.51	(0.08)	0.25	(0.05)	0.34	(0.06)	53	107	(74)	25553
A018	565	(8)	320	(106)	145	(702)	0.80	(0.01)	0.64	(0.07)	0.71	(0.02)	710	885	(114)	30077
H001	0	(0)	108	(122)	0	(0)	NaN	(NaN)	0.00	(0.00)	0.00	(0.00)	0	108	(122)	24336
H002	14	(2)	38	(100)	34	(46)	0.29	(0.04)	0.27	(0.02)	0.28	(0.03)	48	52	(102)	23986
H003	19	(9)	102	(205)	38	(48)	0.33	(0.16)	0.16	(0.04)	0.21	(0.07)	57	121	(214)	24841
H004	0	(0)	23	(100)	4	(4)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	4	23	(100)	26360
H005	2	(3)	50	(133)	4	(3)	0.33	(0.50)	0.04	(0.02)	0.07	(0.04)	6	52	(136)	26362
H006	38	(8)	171	(134)	13	(43)	0.75	(0.16)	0.18	(0.06)	0.29	(0.08)	51	209	(142)	23814
Total	11045	(1001)	4515	(4125)	5732	(15776)	0.66	(0.06)	0.71	(0.20)	0.68	(0.09)	16777	15560	(5126)	605921

in [15]. The authors fully specified their algorithm which enabled its execution on the new data set. It works by extracting two features (peak frequency and 1-lag autocorrelation) from consecutive 3-s windows of a 1-D composite signal (1-s overlap) and then making a decision with a trained logistic regression classifier. As it was designed to be run on single-wrist data at 40 Hz we re-sampled the raw data from 100 to 40 Hz and executed it on both recorded wrists separately. A TP was counted if scratching was assigned in any of the two wrists.

Additionally, in order to evaluate the effect of our integrated algorithm (segmentation based on data characteristics and subsequent RNN classification), we trained and applied both-wrist RNNs, however without the segmentation step described in Subsection II-B. For the RNN-only analysis we subdivided the data from both wrists in consecutive 3-s windows with 1-s overlap.

Based on the objectives listed above the following measures were derived to quantify the algorithms performance.

1) Comparison of Total Scratching Duration: In order to determine the total scratching duration, the amount of 1-s epochs scored as scratching was counted for each measurement method (video, proposed method, RNN-only method and Petersen *et al.* [15]) per recording. To measure general deviation, the median of the differences $\text{median}_{r=1}^{24}(dur_r^{\text{act}} - dur_r^{\text{vid}})$ was calculated, where $r = 1 \dots 24$ denotes the 24 subjects, dur_r^{act} denotes the total duration of scratching for subject r determined using the proposed method or RNN-only method or Petersen *et al.* method and dur_r^{vid} denotes the total duration of scratching for subject r determined by video scoring. Additionally, the spearman-rank correlation coefficient was calculated between the total scratching durations of each subject measured by video and the proposed method, RNN-only method and Petersen *et al.*

2) Comparison of Separability: In order to measure separability between AD and HC groups, the scratching rate was determined (total scratching duration as defined above divided by the recording length). For each measurement method (video, proposed method, RNN-only method and Petersen *et al.*) the receiver operating characteristic (ROC) was calculated (by varying a threshold separating the two groups and plotting the true-positive rate versus the false-positive rate). The area-under-the-curve (AUC) of the ROC can be interpreted as the probability that a classifier will rank a randomly chosen AD subject higher than a randomly chosen HC subject and thus is used as a measure of classification performance [35].

3) Comparison of Temporal Distribution: Each 1-s epoch in a given recording, ignoring those where no scratching was scored in both scorings, was counted either as true positive (TP), false negative (FN) or false positive (FP) independent of the wrist. Based on those counts, sensitivity and precision were calculated.

$$\text{sensitivity} = \frac{TP}{TP + FN} \qquad \text{precision} = \frac{TP}{TP + FP}$$

Both of these measures are summarized in the F_1 score, which is a harmonic mean of sensitivity and precision and thus a measure of overall accuracy. It is determined as

$$F_1 = \frac{2\,TP}{2\,TP + FP + FN}. \tag{5}$$

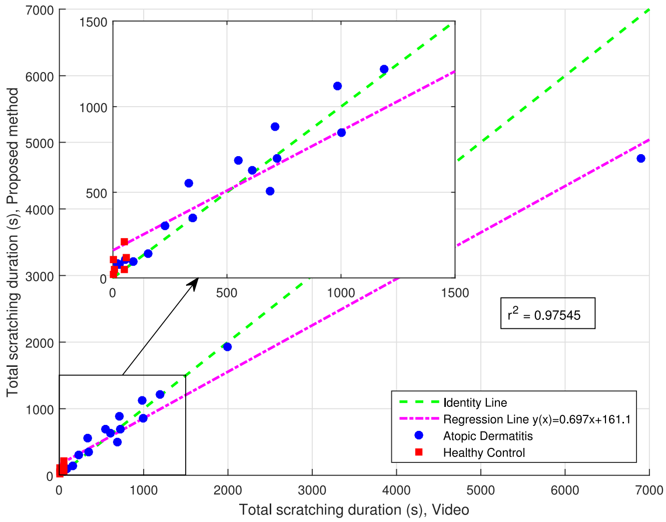

Fig. 4. Scatter plot showing the total scratching duration (video versus the proposed method). The AD subjects are shown in blue circles, the HC in red squares. Due to the outlier recording A013, the majority of recordings is confined in the lower left corner, therefore this section is again shown as an inset enlarged in the upper left corner. In addition, the identity and regression lines are shown, along with the coefficient of determination r^2 as a goodness-of-fit measure.

To determine a measure over all recordings, the number of TP, FN and FP are summed and sensitivity, precision and F_1 score are determined as defined above based on the summed counts.

III. Results

Table I shows the results for the 6-fold cross-validation per subject both for the proposed and the Petersen *et al.* method. The models applied to a given subject were trained on data from other subjects as described in Subsection II-E. The TP, FP and FN counts (in s) are summarized in the Sensitivity, Precision and F_1 score columns. The scratching duration (in s) column lists the amount of scratching scored in a given recording. It is additionally shown in a scatter plot in Fig. 4. The total duration denotes the duration (in s) of the analyzed rest period. Furthermore we report the equivalent results for the RNN only method: 0.93 (total sensitivity), 0.41 (total precision) and 0.57 (total F_1 score).

The deviation of total scratching duration from the gold-standard (video scoring) as determined by Spearman-Rank-Correlation coefficient was 0.945 for the proposed method, 0.914 for the RNN-only method and 0.466 for the method by Petersen *et al.* The median of differences was +37 s for the proposed method, +682 s for the RNN-only method and −112 s for the method by Petersen *et al.* Additionally, the ROC curves that reflect the ability of the scratching rate endpoint to separate between AD and HC subjects are shown in Fig. 5 along with the area-under-the-curve values.

As 12 models were trained (2 per fold) for the proposed method, their performance was evaluated on the independent test set (see Subsection II-E) in terms of error rates (number of wrongly classified sequences divided by total number of sequences in percent). In Tables II and III the error rates are listed per fold.

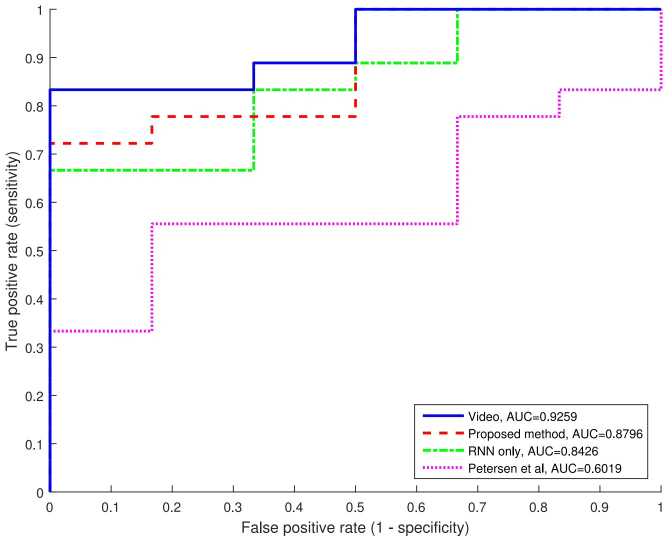

Fig. 5. Receiver Operating Characteristic (ROC) curve that illustrates the ability of the four scoring methods (video, proposed method, RNN only and the method by Petersen *et al.*) to discriminate atopic dermatitis patients from healthy controls. The corresponding ROC area-under-the-curve values are: 0.9259, 0.8796, 0.8426 and 0.6019, respectively.

TABLE II
Error Rates on Data Sets for Single-Hand Models (By Fold)

	1	2	3	4	5	6
training	26.09	29.81	30.57	29.37	20.24	29.32
validation	26.12	15.94	21.79	18.60	30.48	21.38
test	28.44	30.16	30.08	28.77	**27.21**	28.36

TABLE III
Error Rates on Data Sets for Both-Hands Models (By Fold)

	1	2	3	4	5	6
training	21.09	20.19	23.84	20.34	17.27	22.47
validation	9.82	19.42	19.51	12.82	25.26	20.02
test	**19.18**	20.08	22.99	20.48	21.49	19.98

IV. Discussion

The results of the validation study prove that the goal of developing a detection algorithm for nocturnal scratching movements from actigraphy data that shows good performance when compared to the IR video method was met.

The total scratching duration as determined by the gold-standard and the proposed method had a rank correlation of 0.945, with a median of differences of +37 s. Those results suggest good agreement, with a slight over-estimation of scratching in the proposed method. Table I, on the other hand (16777 s for video versus 15560 s for actigraphy), shows a mean difference of −51 s. The discrepancy, however, was entirely explained by subject A013. This subject exhibited 3.4 times more scratching than the subject with the second-most scored scratching.

The F_1 score of 0.68 as compared to 0.09 for the Petersen *et al.* method also demonstrates that the algorithm not only

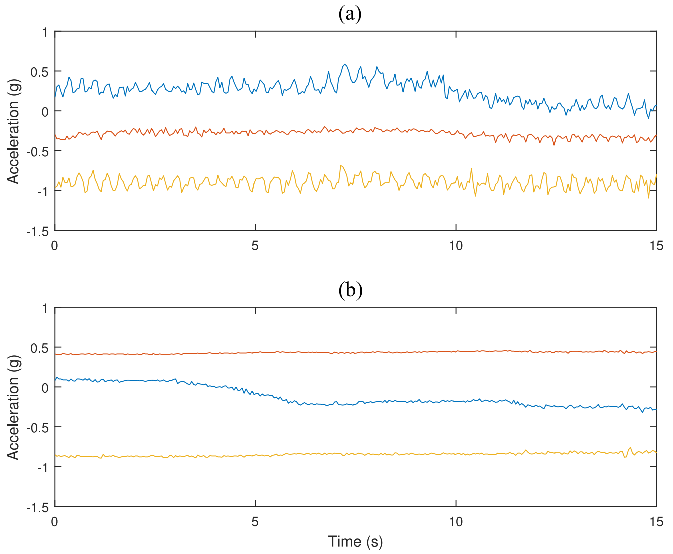

Fig. 6. Two excerpts of left-handed scratching movements: (a) A 15 s excerpt out of a pronounced 59 s scratching movement (signals from top to bottom: x, y, z) where subject A001 scratched the whole length of his leg. (b) A 15 s excerpt out of a subtle 174 s scratching movement (signals from top to bottom: y, x, z) where subject A013 uses only her finger.

correctly estimated the total amount of scratching throughout a recording, but also accurately predicts when the scratching movements were occurring, which could be helpful for determining when during the night a treatment is most effective. While the performance of the Petersen *et al.* method [15] was reported to be quite remarkable (sensitivity $= 0.955$) on their published dataset, the sensitivity drops dramatically to 0.06 (precision $= 0.20$, $F_1 = 0.09$) on the present data set. There are several hypothetical explanations for this discrepancy: Firstly, simulated scratching movements may exhibit different characteristics than spontaneous scratching that might make them easier to discriminate from other movements. Specifically large-area and rhythmical scratching shows quite pronounced features such as periodicity (as can be seen in Fig. 6(a)); it is unclear which types of movements healthy subjects tend to produce when asked to simulate scratching. Secondly, the devices used to record scratching movements were different (the PAM-RL device used by Petersen *et al.* was specifically designed to capture periodic leg movements). Furthermore, the objective of the method by Petersen *et al.* was different: to discriminate defined 30 s excerpts containing scratching from 20 to 30 s excerpts containing restless sleep or walking movements as opposed to detecting scratching movements in a continuous recording.

Scratching movements can be carried out in manifold ways. Large-area scratching involving arm movement (see Fig. 6(a)) can clearly be picked up by the wrist-worn accelerometer sensors, whereas subtle movements of just one finger might not (see Fig. 6(b)). In our method, the th_{noise} threshold was used to detect motion and if its power is so small that it is completely buried in the background noise, the algorithm will not get to classify that particular movement. These are the instances that our algorithm is "blind" to because of the underlying limitations of the sensor placement. In the entire data set (excluding A013) on average about 3% of the total video scored scratching time was affected (which puts an upper limit to sensitivity of about 0.97). The clear outlier recording A013 exhibited large amounts of this kind of subtle scratching (often ranging up to continuous 5 min, see Fig. 6(b)). Here, an unusually high amount of about 13% of scored scratching happened below the th_{noise} threshold. This recording demonstrates the limitations of the proposed method in detecting rather subtle scratching events with little to no arm movements. However, the clinical significance of this type of scratching movement is an open research question.

Tables II and III demonstrate that the models generalized very well, which can be seen on the small differences of error rates on the test set across folds. Nevertheless, the models exhibiting the lowest error rates (27.21 for the single hand model and 19.18 for the both hands model) were chosen to be incorporated in the final algorithm.

The comparison between the integrated approach and RNN-only confirmed the value of our pre-processing segmentation step for identifying scratching candidates. Note that these candidates are characterized not only by increased activity, but also stable wrist orientation. Thus, a major change in the wrist orientation (based on the evaluation of the gravity acceleration vector) terminates a scratching candidate. Indeed our analysis comfirmed an improvement in all performance measures: from 0.57 to 0.68 (F_1 score), from 0.914 to 0.945 (Spearman-Rank-Correlation), from 682 s to 37 s (median of differences) and from 0.8426 to 0.8796 (ROC area-under-the-curve).

A. Limitations and Future Directions

This study has several limitations: The efficacy of the presented algorithm was evaluated on a limited set of subjects, with a low number of healthy controls. This is due to the effort of data acquisition, specifically the manual scoring of IR video. Therefore a replication study with a larger set of subjects including patients with scratching events and healthy controls as well as patients with movement disorders during sleep such as periodic limb movement disorder is advisable. A further limitation is the sensor placement as discussed in Section IV, an alternative sensor placement might be explored.

V. Conclusion

A novel algorithm to detect nocturnal scratching from accelerometer signals was presented. The proposed algorithm was validated on video data from 24 subjects and produced results comparable to the gold-standard video scoring, which demonstrates its effectiveness. It enables cost effective, unobtrusive and longitudinal data collection using wrist worn actigraphy devices to objectively quantify nocturnal scratching with proven accuracy in a clinical setting.

Acknowledgment

The authors would like to thank the Science37 team (Noah Craft, Juan Torres and Katy Preciado) for conducting the study. We also thank Alex Graves for his numerous scientific

contributions to RNN models with LSTM architecture and publishing the RNNLib software.

References

[1] T. Ebata, H. Aizawa, R. Kamide, and M. Niimura, “The characteristics of nocturnal scratching in adults with atopic dermatitis,” *Brit. J. Dermatol.*, vol. 141, pp. 82–86, 1999.

[2] T. Ebata, S. Iwasaki, R. Kamide, and M. Niimura, “Use of a wrist activity monitor for the measurement of nocturnal scratching in patients with atopic dermatitis,” *Brit. J. Dermatol.*, vol. 144, no. 2, pp. 305–309, 2001.

[3] S. Ständer *et al.*, “Pruritus assessment in clinical trials: Consensus recommendations from the international forum for the study of itch (IFSI) special interest group scoring itch in clinical trials,” *Acta Dermato-Venereologica*, vol. 93, no. 5, pp. 509–514, 2013.

[4] M. P. Pereira and S. Ständer, “Assessment of severity and burden of pruritus,” *Allergol. Int.*, vol. 66, pp. 3–7, 2016.

[5] C. S. Murray and J. L. Rees, “Are subjective accounts of itch to be relied on? The lack of relation between visual analogue itch scores and actigraphic measures of scratch,” *Acta Dermato-Venereologica*, vol. 91, no. 1, pp. 18–23, 2011.

[6] Y. Kurihara, T. Kaburagi, and K. Watanabe, “Development of a non-contact sensing method for scratching activity measurement,” *IEEE Sensors J.*, vol. 13, no. 9, pp. 3325–3330, Sep. 2013.

[7] T. Okuyama, K. Hatakeyama, and M. Tanaka, “Frequency response of polymer sensor for measuring finger scratching motion,” *J. Jpn. Soc. Appl. Electromagn. Mech.*, vol. 23, no. 3, pp. 618–623, 2015.

[8] Y. Noro *et al.*, “Novel acoustic evaluation system for scratching behavior in itching dermatitis: Rapid and accurate analysis for nocturnal scratching of atopic dermatitis patients,” *J. Dermatol.*, vol. 41, no. 3, pp. 233–238, 2014.

[9] B. G. Bender, R. Ballard, B. Canono, J. R. Murphy, and D. Y. Leung, “Disease severity, scratching, and sleep quality in patients with atopic dermatitis,” *J. Amer. Acad. Dermatol.*, vol. 58, no. 3, pp. 415–420, 2008.

[10] Y.-S. Chang *et al.*, “Atopic dermatitis, melatonin, and sleep disturbance,” *Pediatrics*, vol. 134, no. 2, pp. e397–e405, 2014. [Online]. Available: http://pediatrics.aappublications.org/content/134/2/e397

[11] K.-L. E. Hon, M.-C. A. Lam, T.-F. Leung, C.-M. Chow, E. Wong, and A. K. Leung, “Assessing itch in children with atopic dermatitis treated with tacrolimus: Objective versus subjective assessment,” *Adv. Therapy*, vol. 24, no. 1, pp. 23–28, 2007.

[12] C. Bringhurst, K. Waterston, O. Schofield, K. Benjamin, and J. L. Rees, “Measurement of itch using actigraphy in pediatric and adult populations,” *J. Amer. Acad. Dermatol.*, vol. 51, no. 6, pp. 893–898, 2004.

[13] K. Benjamin, K. Waterston, M. Russell, O. Schofield, B. Diffey, and J. L. Rees, “The development of an objective method for measuring scratch in children with atopic dermatitis suitable for clinical use,” *J. Amer. Acad. Dermatol.*, vol. 50, no. 1, pp. 33–40, 2004.

[14] J. Feuerstein, D. Austin, R. Sack, and T. L. Hayes, “Wrist actigraphy for scratch detection in the presence of confounding activities,” in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Boston, MA, USA, Aug. 2011, pp. 3652–3655.

[15] J. Petersen, D. Austin, R. Sack, and T. L. Hayes, “Actigraphy-based scratch detection using logistic regression,” vol. 17, no. 2, pp. 277–283, 2013.

[16] J. Lee, D.-k. Cho, S. Song, S. Kim, E. Im, and J. Kim, “Mobile system design for scratch recognition,” in *Proc. 33rd Annu. ACM Conf. Extended Abstr. Human Factors Comput. Syst.*, 2015, pp. 1567–1572. [Online]. Available: http://doi.acm.org/10.1145/2702613.2732820

[17] A. Price and D. E. Cohen, “Assessment of pruritus in patients with psoriasis and atopic dermatitis: subjective and objective tools,” *Dermatitis*, vol. 25, no. 6, pp. 334–344, 2014.

[18] Z. C. Lipton, J. Berkowitz, and C. Elkan, “A critical review of recurrent neural networks for sequence learning,” arXiv preprint arXiv:1506.00019, 2015.

[19] A. Graves, “Long short-term memory,” in *Supervised Sequence Labelling With Recurrent Neural Networks* (ser. Studies in Computational Intelligence), vol. 385. Berlin, Germany: Springer-Verlag, 2012, pp. 37–45.

[20] A. Graves, “Supervised sequence labelling with recurrent neural networks,” Ph.D. dissertation, Technische Universität MünchenGermany, Jul. 2008. [Online]. Available: http://www6.in.tum.de/pub/Main/Publications/Graves2008c.pdf

[21] A. Graves, A.-r. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 6645–6649.

[22] G. Lefebvre, S. Berlemont, F. Mamalet, and C. Garcia, “Inertial gesture recognition with BLSTM-RNN,” in *Artificial Neural Networks*. New York, NY, USA: Springer, 2015, pp. 393–410.

[23] N. Y. Hammerla, S. Halloran, and T. Ploetz, “Deep, convolutional, and recurrent models for human activity recognition using wearables,” arXiv preprint arXiv:1604.08880, 2016.

[24] F. J. Ordóñez and D. Roggen, “Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition,” *Sensors*, vol. 16, no. 1, 2016, Art. no. 115.

[25] *GENEActiv Instructions version 1.2*, Activinsights Ltd., Kimbolton, U.K., Mar. 2012.

[26] J. Hanifin and G. Rajka, “Diagnostic features of atopic dermatitis,” *Acta Dermato-Venereologica. Supplementum*, vol. 92, pp. 44–47, 1980.

[27] Y. Bengio, P. Simard, and P. Frasconi, “Learning long-term dependencies with gradient descent is difficult,” vol. 5, no. 2, pp. 157–166, 1994.

[28] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[29] F. Gers, “Long short-term memory in recurrent neural networks,” Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2001. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.6677

[30] N. Japkowicz and S. Stephen, “The class imbalance problem: A systematic study,” *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, Oct. 2002.

[31] P. J. Werbos, “Backpropagation through time: What it does and how to do it,” *Proc. IEEE*, vol. 78, no. 10, pp. 1550–1560, Oct. 1990.

[32] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, pp. 533–536, 1986.

[33] A. Graves, “RNNLIB: A recurrent neural network library for sequence learning problems,” Aug. 2013. [Online]. Available: http://sourceforge.net/projects/rnnl/

[34] A. Graves, “Practical variational inference for neural networks,” in *Adv. Neural Inf. Process. Syst.*, Granada, Spain, Dec. 2011, pp. 2348–2356.

[35] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.