Detecting Chronic Diseases from Sleep-Wake Behaviour and Clinical Features.

Sarah Fallmann
School of Computer Science and Informatics
De Montfort University, Leicester, U.K.

Liming Chen
School of Computer Science and Informatics
De Montfort University, Leicester, U.K.

***Abstract*—Many chronic diseases show evidence of correlations with sleep-wake behaviour, and there is an increasing interest in making use of such correlations for early warning systems. This research presents an approach towards early chronic disease detection by mining sleep-wake measurements using deep learning. Specifically, a Long-Short-Term-Memory network is applied on actigraph data enriched with clinical history of patients. Experiments and analysis are performed targeting detection at an early and advanced disease stage based on different clinical data features. The results show for disease detection an averaged accuracy of 0.62, 0.73, 0.81, 0.77 for hypertension, diabetes, sleep apnea and chronic kidney disease, respectively. Early detection performs with an averaged accuracy of 0.49 for sleep apnea and 0.56 for diabetes. Nevertheless, compared to existing work, our approach shows an improvement in performance and demonstrates that predicting chronic diseases from sleep-wake behavior is feasible, though further investigation will be needed for early prediction.**

***Index Terms*—Deep Learning, Chronic Disease Detection, Sleep Monitoring.**

I. Introduction

Chronic diseases are one of the main causes of deaths and disabilities which is a major factor to the quality of life of patients and relatives. With the growing ageing population, the prevalence of chronic diseases will impact more and more people. In addition, chronic diseases are also a problem for the society as they produce high costs for recovery and management [1]. Currently, chronic diseases are under-diagnosed by a significant percentage, by using the estimated effected patients against the once diagnosed; hypertension $30-40\%$, diabetes $20-50\%$ [2] and it is also present in chronic kidney disease (CKD) [3] and obstructive sleep apnea [4]. An easy, sensor-based and reliable method for early diagnosis or prediction is still missing.

It is clinically observed that people suffering from chronic diseases such as diabetes [5], [6], CKD [7], hypertension [8], arthritis and stroke [6] usually have troubles with their sleep, like difficulty falling/staying asleep and daytime sleepiness [6]. These correlations with sleep-wake behaviour can be used for prediction of chronic diseases, which is investigated in the presented approach. The differences are manifested in (1) the actigraph sleep-wake patterns of the individuals, comprising information of the sleep efficiency and sleep duration [9], and (2) the clinical feature characteristics, such as body mass index(BMI). The diagnosis and early detection of specific chronic conditions and comorbidities can also later help in the self-management process [1].

Diagnosis of chronic disease are currently based on (1) invasive methods such as blood sugar screening for diabetes [10], (2) symptoms, risk factors and clinical history or (3) sensor-based, such as within a sleep clinic to diagnose obtrusive sleep apnea [4] or with blood pressure measurements for hypertension [11]. These methods can also be combined to reach a fast and trustworthy result, for example, by combining medical history, physical exam, and results from a sleep study for OSA diagnosis [4].

For early detection, two main methods are presently used (1) Marker-based clinical analysis and (2) sensor-based behavioural analysis.
In method (1), early detection of chronic diseases can be based on bio-, neuropsychological- or structural-image-markers. Research has been done on, e.g., Alzheimer's disease [12], diabetes [10] and CKD [3].
The second method (2) is based on data investigation coming from sensors, such as an actigraph. Ju *et al*. [13] analysed the role of sleep in early detection of Alzheimer's disease in humans based on β-amyloid(Aβ) and actigraph data. Aβ is related to sleep quality and quantity. For participants with low Aβ42, it showed a correlation with worse sleep quality. Aggarwal *et al*. [14] investigated actigraph data from the Hispanic Community Health Study [15] to detect sleep apnea, insomnia, diabetes and hypertension. They applied a Convolutional Neural Network (CNN) and presented a method for embedding activities. The study has some limitations as such: An imbalanced dataset, used for a supervised machine learning approach, with the majority of subjects not suffering from the disorders and a classification task to predict the disorder-positive subjects. Their models and baselines are highly biased towards predicting the majority class [14]. Actigraph data is also used for exploring correlation of sleep behaviour disorders between affected and non-affected chronic disease patients such as Parkinson's disease. This includes, for example, that the total number of wake bouts is higher for Parkinson disease patients affected by sleep behaviour disorders [16]. In a later stage, this knowledge can potentially be used for early detection of specific disease related symptoms and features,

and help during the diagnosis process. Time-series data based diagnosis, in the field of recognising patterns in multivariate time-series of clinical measurements, is presented in [17], which has been a quite new approach, as previous studies mostly applied neural networks in this context, but not Long-Short-Term-Memory networks (LSTMs).

The paper contributes to the literature in three folds. Firstly, we apply the Long-Short-Term-Memory networks (LSTMs) which, in contrast to previous work, can deal with the temporal aspects of the actigraph data. Secondly, we fuse actigraph and clinical history data of individuals, introducing a multi-dimensional feature vector. Thirdly, we experimented and analysed multiple use scenarios, including the good and bad classified patients based on significant difference of clinical features. Through these novel approach and experimental studies, we address the challenges for chronic disease detection and limitations from previous work.

The paper is structured in the following way. In Section II the approach for the chronic disease recognition is described, including the initial method configuration. The dataset is described and the experimental settings are given. Section III presents the results which are described and discussed. Section IV concludes the detection approach and describes possible future extensions.

II. Method and Materials

In this section, an approach is proposed to diagnose and early detect chronic diseases based on sleep-wake behaviour and clinical history data. Furthermore, the method design and configuration using the time-series classification algorithm are presented. The dataset used for investigating the approach and the settings of the experiments are given.

A. Learning Model Design and Configuration

The approach is a time-series data classification. Time-series data are data which have time dependency, in our case those are activity counts and white light data from six consecutive days. Activity counts give the intensity of activeness over time. White light is an actigraph feature which can tell the level of darkness or lightness at a specific time, helping to judge the day-night-rhythm, further details can be found in Section II-B. A prominent deep learning method in the field of time-series data classification is a deep Neural Network, called Long-Short-Term-Memory networks (LSTMs). This method was first introduced by Hochreiter and Schmidhuber [18] in 1997, extended over time mainly by additions of forget gates [19], and peep-hole connections [20]. Application in the field of diagnosis of multiple time-series data already showed successful [17], following an architecture described by Graves *et al*. [21].

In this paper, a LSTMs approach is followed using the architecture presented in Fig. 1. This contains an embedding based on word-2-vec for the activity count and white light data. Embedding is able to extract useful features to represent similarities and relationships within data. Words are translated into a vector representation [22], [23]. Here, activity counts and white light data are interpreted as words. The embedding method provides features used as input for LSTMs [22], [23]. Clinical history data are further sources of information used as direct input. Clinical history data include clinical knowledge of the person such as BMI level or Family History of Diabetes, details on the features used can be found in Section II-B. The combination of actigraph data enriched with clinical history data from individuals is used as input for three LSTMs layers each using a dropout strategy for training, followed by a dense layer to flatten and fully connect. Dropout is used to ignore neurons during the training phase to prevent overfitting [24]. The fraction of dropout can be set between 0.0 (no dropout) to 0.9. For example setting the dropout rate to 20% means that two in ten inputs will be randomly excluded. Dense layers apply a matrix-vector multiplication resulting in a transformation of the data into a lower dimensional feature vector. The last step is a softmax activation, which normalises scores to probabilities for classification of two or three classes, whereas two classes are (a) disease affected and (b) non-specific-disease-affected and 3 classes are (a) pre-disease-, (b) disease- and (c) non-specific-disease-affected. The two classes problem implements the disease detection distinguishing affected and non-affected individuals whereas the three classes problem deals with early disease detection distinguishing affected, early stage and non-affected.

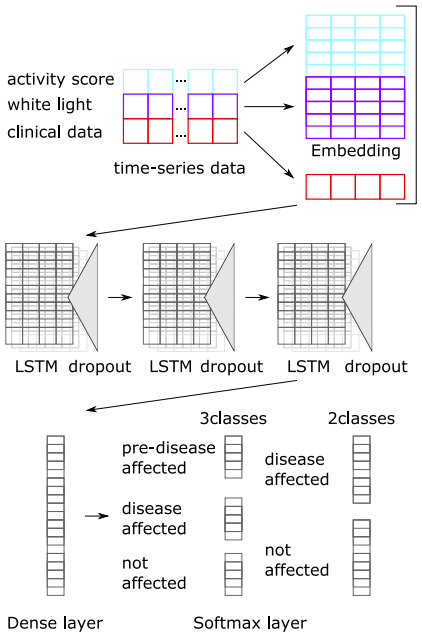


Fig. 1: The LSTMs method architecture for disease detection.

Initial experiments let us choose the following layer sizes: 1^{st} layer input size is [400+clinical data features] and output size 200, 2^{nd} layer input 200 and output 100, and input of

100 and output of 50 for the 3^{rd} layer. Furthermore, the combination of a dropout rate of 0.2 and batch-size allowing training per person is used, compare Fig. 4 and Section II-C. The model fitting was investigated with an adam optimisation method [25] as initial experiments ruled out Root Mean Square Propagation for our study. Embedding size was empirically set to 200, where only the activity levels and white light data are embedded. This makes it possible to represent the data sequence (e.g. 240) in a feature vector [23] with size 200, creating an input vector for the LSTMs of size 400+clinical data features, consisting of 200 features for activity counts, 200 features for white light and the clinical data features different for each disease, see details in Section II-B. The Training Process was stopped when the loss did not decrease within 8 intervals, with a maximum of 50 epochs.

B. Dataset and Processing

The data used to show the possible improvement by using the established approach, was made available by the National Sleep Research Resource [26] and includes clinical history and actigraph measurements from 2,252 participants with different disorders such as hypertension, apnea, diabetes and CKD. Clinical history includes information about an individual such as BMI level or Family History of Diabetes. The participants were instructed to wear wrist-worn actigraph devices (Actiwatch Spectrum, Philips Respironics) for a week. Records were scored by a trained technician at the Boston Sleep Reading Center [26] and specific algorithms, deriving parameters such as wake/sleep patterns and activity levels. Actigraph data are available per 30 seconds, which make 2880 values per day. The data of six consecutive days per person are used for analysis in this study, as not all participants have seven full days of data. This is necessary to provide a balanced dataset for training purposes. Initial results show that using the same amount of data for each individual results in better outcomes, which was also the case for Aggarwal *et al.* [14].

The proof of concept is investigated for four different chronic diseases and conditions: (1) diabetes, (2) hypertension, (3) CKD and (4) sleep apnea. These are divided into three different classes each, which are (a) non-specific-disease-affected (b) pre-disease (early stage) and (c) disease-affected, compare Fig. 2(a). For hypertension only two classes are available.

(1) Diabetes is characterised by hyperglycemia [27]. Prediabetes is classified over the elevation of plasma glucose above the normal range but below that of clinical diabetes [27]. People affected by diabetes are more likely to have trouble sleeping, caused by the glucose level which is influencing sleep. Furthermore, lethargy and insomnia can be caused by diabetes [28].
(2) Hypertension is defined by an abnormal blood pressure ($\geq$ 140/90) [11]. Sleep and hypertension are correlated, in more detail, habitual short sleep duration is associated with hypertension. Insomnia is correlated with increased hypertension risk [8].

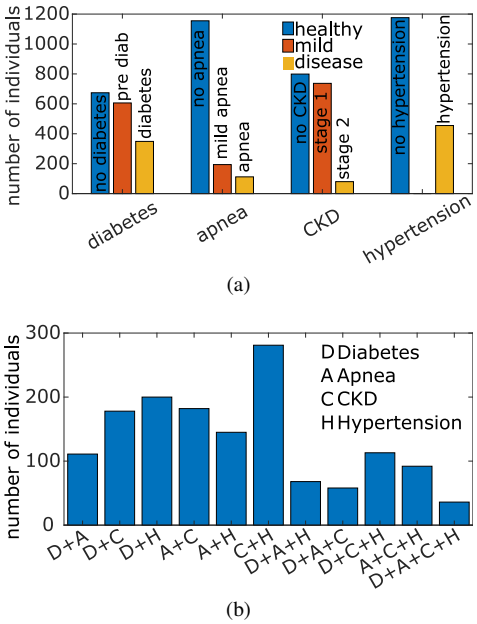


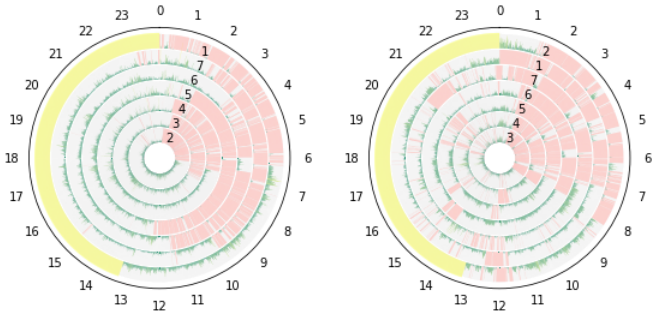
Fig. 2: (a) The number of people per disease divided by not-specific-disease affected, mildly affected and affected. (b) Multi-morbidities in the dataset based on diabetes, apnea, CKD and hypertension.

(3) CKD is a severe health problem, which is defined as kidney damage over glomerular filtration rate (GFR). Five stage classification is based on the level of GFR [3]. In this experiments, only stage one and two are investigated. A wide range of symptoms affects the patients leading to e.g., fatigue, sleep disturbances, restless legs and increase in urination mostly during the night [7].
(4) Obstructive sleep apnea is classified by upper airways becoming blocked repeatedly during sleep [4]. The classification is based on the Apnea/Hypopnea Index (AHI), where the numeric events per hour are given. In the concrete use-case, >4% oxyhemoglobin desaturation is used for the AHI. These can occur less than 5 times per hour resulting in no apnea, between 5 and 15 times per hour classified as mild apnea, between 15 to 30 times resulting in moderate apnea and severe apnea with occurrences over 30 times per hour [4]. In the test case three groups where used, which are none, mild and moderate-severe.

In Fig. 2(b) the number of people in the dataset with more than one disease and the combination of which is given. Showing combination of diseases like apnea and diabetes, diabetes and hypertension and others. Different multi-morbidities are shown to be more present than others and therefore can influence the classification, as the non-specific-disease-affected classified people, can potentially be affected by another disease.

To give an overview of the data and the individual differences, between affected and non-affected patients a circle plot is given for diabetes, compare Fig. 3. The plots show activity count data (green) over the intervals of being awake (grey) and

asleep (red) over 7 consecutive days. The yellow describes not available data due to differences in the end times. The plots illustrate non-diabetic-affected Fig. 3(a) and diabetes-affected Fig. 3(b). The images show the already affected and not mild or pre-stage affected individuals. The images should be seen as an example of individuals classified correctly by the algorithm. Be aware of, that there are personal differences between individuals as well, but general conclusions can be drawn for specific disease groups, as the results show in Section III. The use of disease-affected and non-specific-disease-affected is a real world approach. It needs to be mentioned that this can result in a group which are not affected by the specific disease, but can also not be classified healthy, as they can for example not be affected by apnea, but still be affected by diabetes, as multi-morbidities are common, compare Fig. 2(b).



(a) Not-affected by diabetes (b) Affected by diabetes

Fig. 3: Acitivty count data of seven days from individuals, affected by diabetes and non-affected, shown in green and showing the wake and sleep patterns in red and grey.

Preprocessing includes extraction of individuals and data of interest coming from (1) actigraph data, including activity levels and white light and (2) clinical data, including BMI, family history of diabetes and affected by hypertension, diabetes or asthma. Activity levels and white light data are used for all disease-specific tasks and specific clinical data parameters are used for each disease. For CKD, risk factors contain diabetes, high blood pressure as well as obesity [29] and therefore are included as clinical history information (BMI, diabetes and hypertension). For apnea, comorbidities usually are obesity, diabetes and hypertension and sometimes asthma [30], therefore those are included (BMI, asthma, diabetes, hypertension). For diabetes the family history of diabetes is important and obesity [31] (BMI, diabetes family history). Hypertension is affected by obesity and is a risk factor for diabetes [32], [33] so correlated and included in the clinical history information (BMI, diabetes).
Furthermore, the data is normalised for training and testing during the classification process.

C. Experiment Design and Implementation

In this paper, a LSTMs approach is followed and compared based on training and test performance in different experiments and compared with an existing method from [14] which used CNN. The main improvement is reached by introducing disease-specific features coming from clinical data history based on important relations from literature, compare Section II-B.

LSTMs parameters were adjusted empirically by initial experiments, details can be found in Section II-A. To reach training per person the batch-size for training and testing is six, as six days per individual are available. Splitting the daily data from 2880 sequence length (dimension for one day 2880x1) into 180 (dimension for one day 180x16) or 240 (240x12) implies a batch-size training rate of 96 (16*6) or 72 (12*6), respectively, compare Fig. 4. This guarantees a training per person approach which means 6 days per person. Sequence length of 2880(one day) was not further used, based on the long training time and no increase in accuracy.

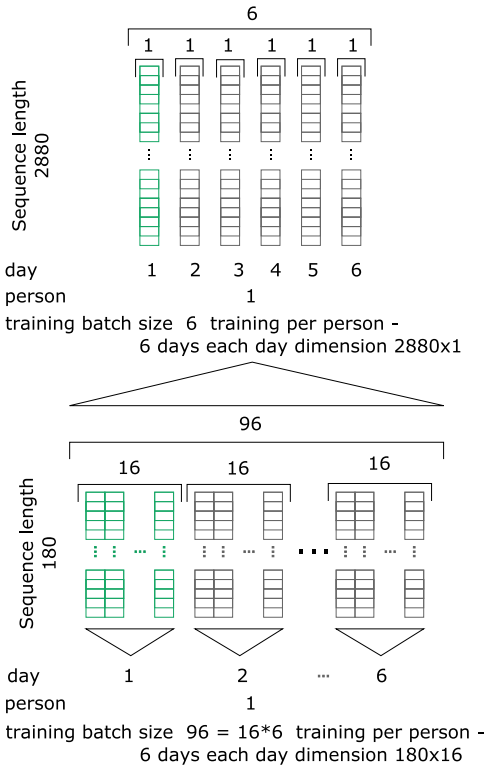


Fig. 4: Training batch size transformation for the example of 180 sequence length.

The first experiment investigated the best sequence length. The sequence lengths tested are 180 and 240 for each disease-specific classification for 2 and 3 classes.

The repeated wrong classification of certain participants suggests investigating the wrong classified participants in more detail to extract differences in those. Those differences can be seen as potential sources for misclassification which can be used as additional information for improving the classification approach. The statistic counts for non-specific-disease-affected wrong/correct-classified and disease-affected wrong/correct-

classified are given. Wrongly classified are those which are more than two times wrong classified during testing. The investigation concentrates on the best model outcome. The second experiment classifies the significant different clinical data parameters from the wrong and correct classified disease-affected individuals during the test. The clinical data was investigated in more detail with an independent two samples t-test, for different means, variances, and sample size. Here, the independent variable is disease-affected correct-classified from disease-affected wrong-classified, and the dependent variable is one of 38 clinical data parameters pre-decided from the available variables in the dataset [26], [34]. Using a significance level of 5%, different number of clinical data parameters can be extracted being significantly different from the disease-affected correctly classified from those wrongly classified. These are given with the p-value of the statistical test. Those clinical parameters include among others BMI, Family History of Diabetes and Hypertension.

The third experiment investigated the possible improvement which can be reached by including the knowledge of having 6 days of data per person. This is done in two ways based on (1) counts and (2) average. The data per person as for sequence length 180 is 96. (1) counts how often which class is classified between e.g. intervals of length 96 and the one with the highest amount, or the first one in the list if there are two, is reassigned to the whole interval to 96 values. (2) takes the average of the interval using a threshold of 0.5.

Furthermore, the outcomes of the best experiment, based on the sequence length experiment, are compared with available related work results from [14]. Task-specific means the models are trained end-to-end for the disorder task at hand, whereas, multi-task learning models, are trained jointly with all the disorder prediction tasks learnt jointly (in the case of [14] those are insomnia, apnea, diabetes and hypertension).

The experiment results show the averaged outcome for each experiment which follows a leave-n-people-out approach. The value n is 10% of the available disease affected individuals for each disease-specific classification: diabetes, hypertension, apnea and CKD. This means a 10-fold-cross-validation for participants is conducted, with training on 90% of the data and testing on 10%. Furthermore, each experiment is repeated six times. These are 60 validation performances. Per fold the best 'area under the curve(AUC)' and best 'accuracy' of all epochs are chosen to give a 10-fold average outcome, which is averaged over the six repeats. The reported outcomes are accuracy, precision, recall, F_1-score and AUC.

Experiments are the following: (1) Disease detection based on a two classes problem for apnea, CKD, hypertension and diabetes, which detects disease affected by using different sequence lengths. (2) Early disease detection for diabetes and apnea, which detects early stage and disease affected resulting in a 3 classes classification problem, testing different sequence lengths. (3) Significant clinical data features for wrong and correct classified disease-affected individuals. (4) Improvement of accuracy by including knowledge of six days of data per person and (5) Comparison of our results to available related work results.

III. Results and Discussions

In this section, the results of the different experiments are presented for disease and early disease detection. Furthermore, results are described, discussed and limitations are examined.

A. Disease Detection

In summary, the disease detection shows overall good outcomes for diabetes, apnea and CKD. For hypertension the general outcome is less promising, see Table Ib and IV, compared to the results in [14] no improvement in accuracy can be seen, but in precision, recall and F_1-score. For apnea detection our approach can outperform the results with CNN, compare Table IV. The results of the 2 classes experiments are compared to the results from [14] for hypertension and apnea.

Diabetes Experiments on averaged AUC and accuracy show that for the case of diabetes similar results can be achieved with sequence length of 240 and 180 resulting in an AUC of $0.76(\pm 0.02)$ and accuracy of 0.71 for 240. A higher standard deviation can be seen for 240 in comparison to the 180 case for averaged accuracy, compare Table Ia. Therefore, the best outcome is classified as the 180 sequence length case for averaged accuracy.

For diabetes detection, 120 disease-affected individuals are wrongly classified as non-specific-disease-affected over all tests and 205 are correctly classified as disease-affected. For the non-specific-disease-affected participants, 107 are wrongly classified from randomly picked 325 individuals from all non-specific-disease-affected participants in the dataset. In Table II three variables have been shown to be significantly different in the disease-affected case comparing correct classified and wrong classified. Those are BMI, Family History of Diabetes and Pneumonia. These features are affecting the wrong classified features. The group that is wrongly classified has apparently (1) a higher average BMI with 33.3 than the group that is usually correctly classified with 31.6, (2) the average family history of diabetes is 0.9 in contrast to 0.5 for correctly classified and (3) for pneumonia the average of appearance is 0.22 and 0.12 for correctly classified.

The results can be improved when the knowledge of training and testing per person is used. This means the data structure, of 6 days of data per person. This results in an increase of accuracy by around 1% to 72.5 and 72.7 percent for 180 and 240, respectively, making the combined scores the best optimisation technique for this case, compare Table III.

Hypertension Comparison with the already performed experiments of [14] in Table IV shows that the LSTMs approach improves the precision, recall and F_1-score, but can reach

TABLE I: Experiments for different sequence lengths are given for the tested chronic diseases. Each table show the averaged cross-validation outcome for precision(P), recall(R), F_1-Score, Area under the curve(AUC) with Standard deviation(SD) and accuracy(acc), for averaged accuracy and averaged AUC, compare Section II-C.

(a) Diabetes

size	avg	P	R	F_1	AUC(SD)	acc
180	acc	0.70	0.77	0.73	0.75(±7E-3)	0.71
180	AUC	0.69	0.74	0.70	0.76(±7E-3)	0.69
240	acc	0.69	0.79	0.73	0.76(±0.02)	0.71
240	AUC	0.67	0.77	0.72	0.77(±0.02)	0.69

(b) Hypertension

size	avg	P	R	F_1	AUC(SD)	acc
180	acc	0.60	0.64	0.61	0.60(±0.01)	0.60
180	AUC	0.58	0.60	0.58	0.60(±0.01)	0.58
240	acc	0.58	0.60	0.58	0.56(±2E-3)	0.58
240	AUC	0.55	0.57	0.55	0.57(±2E-3)	0.55

(c) Apnea

size	avg	P	R	F_1	AUC(SD)	acc
180	acc	0.82	0.77	0.78	0.82(±0.05)	0.79
180	AUC	0.79	0.75	0.76	0.85(±0.05)	0.77
240	acc	0.83	0.75	0.78	0.81(±0.07)	0.79
240	AUC	0.81	0.70	0.74	0.84(±0.07)	0.76

(d) Chronic Kidney Disease

size	avg	P	R	F_1	AUC(SD)	acc
180	acc	0.80	0.71	0.73	0.72(±0.15)	0.75
180	AUC	0.75	0.63	0.66	0.78(±0.15)	0.69
240	acc	0.79	0.62	0.67	0.68(±0.07)	0.71
240	AUC	0.74	0.55	0.59	0.73(±0.07)	0.66

TABLE II: Significant parameters analysis outcome of the independent 2-sample t-test are given for each disease, based on the disease-affected wrongly and correctly classified individuals.

disease	parameters(p-value)	
diabetes	BMI(0.02) Pneumonia(0.018)	Fam. Hist. Diab.(2.1E-09)
hypert.	BMI(1.8E-17) Apnea/Hypopnea Index(4.8E-05) Alcohol Use (0.028)	Total Drinks per Week(0.012) Treatment of Hypert.(0.013) Physical Activity Level(0.036)
apnea	BMI(2e-3) Hypertension(1E-3) Diabetes Diag.(0.025) Antihypertensives(0.025) Hypertension Treatment(0.015)	4-level Hypertension(3E-3) 3-level Diabetes(0.04) Antidiabetics(0.036) Cigarette Use(0.017)
CKD	BMI(0.032) Hypertension(1.4E-3) Airflow Limitation(0.044) 4-level Hypert.(0.021)	Antihypertensives(0.025) Treatment of Hypert.(0.012) Physical Activity Level(0.043) Stroke(0.036)

TABLE III: Accuracies using the knowledge of 6 days per person. Each table shows the accuracy(acc), the accuracy calculated by using the highest count of labels per person(counts) and the best score per person(average), details can be found in Section II-C.

disease	size	acc	counts	average
diabetes	180	0.714	0.725	0.725
	240	0.712	0.726	0.727
hypert.	180	0.603	0.615	0.620
	240	0.578	0.587	0.589
apnea	180	0.794	0.813	0.810
	240	0.791	0.808	0.800
CKD	180	0.750	0.756	0.767
	240	0.709	0.728	0.722

only similar accuracy for sequence length 180 for the multi-task case. These results can come from the fact that the used

TABLE IV: Comparison Task-specific and Multi-task from [14] with our results from apnea, hypertension and diabetes.

experiment	model	precision	recall	F_1	acc
HYPERTENSION					
task-spec	CNN	0.44	0.31	0.37	0.69
multi-task	CNN	0.48	0.41	0.44	0.61
180	LSTM	0.60	0.64	0.61	0.60
240	LSTM	0.58	0.60	0.58	0.58
APNEA					
task-spec	CNN	0.31	0.63	0.42	0.55
multi-task	CNN	0.40	0.48	0.43	0.68
180	LSTM	0.82	0.77	0.78	0.79
240	LSTM	0.83	0.75	0.78	0.79
DIABETES					
task-spec	CNN	0.40	0.51	0.45	0.41
multi-task	CNN	0.46	0.47	0.46	0.44
180	LSTM	0.44	0.58	0.50	0.47
240	LSTM	0.44	0.57	0.51	0.48

embedding is not designed for activities, but for words.

Overall, we can see that hypertension is the least well performing classification in the 2-class-problem, compare Table Ib. The best outcome is achieved with 180 with an accuracy and AUC of 0.60, with very little variation, which suggests that specific groups cannot be detected. This suggests investigating the wrong classified participants in more detail to extract differences in those. The investigation showed that those which are more than two times wrong classified with sequence length 180 is 143 over all tests and 332 being correct classified as disease-affected. For the non-specific-disease-affected participants, 126 are wrong classified from randomly picked 475 used from all non-specific-disease-affected participants. Using a significance level of 5%, six variables have been shown to

be significant different between the disease affected correct classified from those wrong classified. Those are BMI, Total drinks per week, AHI, Treatment of Hypertension, Alcohol Use and Physical activity level, compare Table II. These features are affecting the wrong classified features. The group that is wrongly classified has apparently differences in the average of the features: (1) higher BMI 34 than correctly classified with 30, (2) lower Total drinks per week 1.28 compared to 2.56, (3) higher AHI 7.5 compared to 3.9 (4) higher Treatment of Hypertension 0.62 compared to 0.5 (5) lower Alcohol Use 2.05 compared to 2.23 and (6) higher Physical activity level 2.52 compared to 2.4.

An improvement of 1-2% in accuracy can be seen when using the combined scores method for the accuracy, resulting in $0.59(240)$ and $0.62(180)$ accuracy, see Table III.

Sleep Apnea The best outcome for sleep apnea can be reached using averaged accuracy resulting in a similar outcome for sequence length of 180 and 240, compare Table Ic. An accuracy of 0.79 can be reached and an AUC of 0.82. The investigation showed that the wrongly classified as non-specific-disease-affected are 23 and 62 being correctly classified as disease-affected. For the non-specific-disease-affected individuals, 13 are wrong classified from 85 randomly picked from all non-disease participants. Those wrong/correct classified from the disease affected group have a significant difference in the following variables: BMI, 4-levels Hypertension, Hypertension, Hypertension Treatment, 3-levels Diabetes, Diabetes Diagnosis, Antidiabetics, Antihypertensives and Cigarette Use, see Table II. The group that is wrongly classified has differences in the average of the features: (1) higher averaged BMI of 38 than correctly classified with 33, (2) higher 4-level grouped Hypertension 3 compared to 2.3, (3) higher Hypertension 0.8 compared to 0.4 (4) higher 3-level grouped Diabetes 2.3 compared to 2.0 (5) lower Diabetes Diagnosis 2.2 compared to 2.6, (6) lower Antidiabetics 0.27 compared to 0.07, (7) higher Antihypertensives 0.47 compared to 0.2, (8) higher Cigarette Use 1.9 compared to 1.5 and (9) higher Hypertension Treatment 0.6 compared to 0.3.

An improvement of accuracy is depicted in Table III with the combined amount technique reaching 0.813 and 0.808 for 180 and 240, respectively. In comparison, the results of [14] in Table IV can be improved considerably by using LSTMs from an accuracy 0.68 (CNN) to 0.79 (LSTM) and precision even from 0.40 (CNN) to $0.82(180)$, $0.83(240)$ (LSTM).

Chronic Kidney Disease For CKD detection the best model can be classified as the 180 model with an accuracy of 0.75 and AUC of 0.72, compare Table Id. Receiver Operating Characteristic curves were interpreted and showed a higher standard deviation for the individual True Positive rates, which represents the higher differences in epochs and folds. An improvement for accuracy can be reached for the sequence length 180 with the combine score technique towards 0.767 and for the 240 sequence length with the combined amount reaching 0.728, compare Table III.

The investigation showed that the wrongly classified for sequence length 180 are 17, and 43 being correct classified as disease-affected. For the non-specific-disease-affected participants 11 are wrong classified. 8 variables have been shown to be significant different. Those are BMI, Antihypertensives, Hypertension, Treatment of Hypertension, Airflow limitation, Physical activity level, 4-level grouped hypertension and Prevalent Stroke (self report), compare Table II. Those present features are potentially more than in the other diseases, because of the less available data. Differences in the average for wrong/correct classified are as follows: (1) higher averaged BMI of 35 than correctly classified with 30, (2) higher 4-level grouped hypertension 2.9 compared to 2.3, (3) higher hypertension 0.8 compared to 0.4 (4) lower Airflow limitation 0 compared to 0.1 (5) higher Physical activity level 2.6 compared to 2.3, (6) higher Prevalent Stroke 0.3 compared to 0.03, (7) higher Antihypertensives 0.6 compared to 0.3 and (8) higher Treatment of Hypertension 0.7 compared to 0.3.

B. Early Disease Detection

The early detection experiments show the outcomes of the 3-classes problem, differentiating non-disease, pre-disease and disease affected. As the standard deviation of CKD was already high for the 2-classes problem, we concentrated on diabetes and sleep apnea. The experimental outcomes show the possibility of early detection. The results should be seen as preliminar, which can further be improved and investigated with experiments including further features or other classification models.

Diabetes In Table IV the 3-classes problem of Diabetes is compared to the related work in [14], showing that our approach can reach an improvement of 3-7% in accuracy from 0.41 (CNN: Task-specific), 0.44 (CNN: Multi-task) to 0.47 (LSTM:180), 0.48 (LSTM:240) and also recall and F_1 show an improvement overall. Precision has similar outcomes.
For early detection of diabetes similar outcomes can be achieved for sequence length of 180 and 240, resulting in averaged accuracy of 0.47 (180) and 0.48(240), and averaged AUC of 0.63 (180) and 0.61 (240), compare Table Va. The accuracy can be improved towards 0.48(180) and 0.49(240) with the count method.

Sleep Apnea For sleep apnea the best results can be achieved with sequence length 240 with an averaged accuracy of 0.54 and an averaged AUC of 0.61, compare Table Vb. The accuracy can be improved towards 0.55(180) and 0.56(240) with the count method. The individuals which are repeatedly wrong classified for the pre-disease-affected are 9, disease-affected are 24 and non-specific-disease-affected are 18. Overall 85 individuals per group are used for training. The features which are significantly different for the disease-affected wrong/correct classified are: BMI (p-value:0.012), Antihypertensive (0.028), Hypertension (0.0018), Hypertension Treatment (0.029), 3-level Diabetes (0.049), Diabetes

TABLE V: Sequence lengths experiments for early disease detection with a multi-class analysis approach, using one class against the others; e.g., C1 describes C1 against all other classes; C1 is non-disease, C2 disease-affected, C3 pre-disease; avg. describes the average over all classes and all repeats.

(a) Diabetes.

size	precision				recall				F_1				AUC(Std dev.)				acc
	C1	C2	C3	avg	C1	C2	C3	avg.	C1	C2	C3	avg.	C1	C2	C3	avg.	
180	0.48	0.49	0.35	0.44	0.58	0.59	0.50	0.58	0.53	0.56	0.34	0.50	0.66	0.66	0.51	0.62	0.47
240	0.48	0.50	0.35	0.44	0.58	0.60	0.51	0.57	0.53	0.58	0.35	0.51	0.65	0.67	0.53	0.62	0.48

(b) Apnea.

size	precision				recall				F_1				AUC(Std dev.)				acc
	C1	C2	C3	avg	C1	C2	C3	avg	C1	C2	C3	avg	C1	C2	C3	avg	
180	0.59	0.55	0.37	0.48	0.57	0.56	0.46	0.57	0.58	0.66	0.34	0.56	0.71	0.68	0.49	0.63	0.53
240	0.56	0.49	0.38	0.46	0.56	0.55	0.47	0.56	0.59	0.50	0.30	0.52	0.69	0.64	0.49	0.61	0.54

Diagnosis (0.019), Family History of Diabetes (0.038), Pneumonia (0.017) and 4-level Hypertension (7.6e-4). The group that is wrongly classified disease-affected has differences in the average of the features to the correctly disease-affected classified: (1) higher averaged BMI of 38 compared to 34, (2) higher Antihypertensive 0.5 compared to 0.2, (3) higher Hypertension 0.8 compared to 0.4, (4) higher Hypertension Treatment 0.58 compared to 0.3, (5) higher 3-level Diabetes 2.4 compared to 2.0, (6) higher Family History of Diabetes 0.7 compared to 0.4 and (7) higher Pneumonia 0.3 compared to 0.04 and (8) higher 4-level Hypertension 3.0 compared to 2.3. We can see that similar features are significant in the 2-classes apnea problem.

Overall, we can conclude that the pre-disease condition is harder to classify. The early detection seems to be complex and in need of further insight with more features or data sources. Potential sources could be information from polysomnography and clinical data features being significant different in the wrong/correct classified cases.

C. Limitations

Limitations are the usage of pre-defined embedding function and the missing evaluation of different loss functions. Initial experiments predefined embedding showed good results and was used for further investigation. The used word-2-vec embedding method can be improved by using a designed act-2-vec method as presented in [14]. Based on literature the categorical cross-entropy approach was used for classification, but others were not tested.
The decision was made to use the same amount of data for each individual, as in 6 days each. Due to this decision a limitations arise, as (1) some individuals can have a full weekend in their dataset and some just one day and (2) not all participants have the same consecutive days sequence, which can lead to difficulties during the training phase of the classification algorithm.

The significant feature analysis for the early disease detection could give more insight by grouping the individuals which are wrong-classified into two groups, wrong-classified as (1) non-specific-disease-affected and (2) pre-disease-affected.

IV. Conclusion and Future Work

This paper has investigated the hypothesis of chronic disease detection based on the correlations between sleep-wake behavior and chronic diseases. An LSTM algorithm was applied fusing actigraph data with clinical features, leading to a multi-dimensional feature vector for effective disease detection. We have tested and evaluated various use case scenarios, including 2-classes disease detection problems (non-specific disease versus disease affected) and 3-classes early disease detection problems (non-specific-disease-affected, early stage, and disease-affected), for four typical chronic diseases, as such, hypertension, diabetes, sleep apnea and chronic kidney disease. The approach is showing promising results, varying for different use cases, resulting from existing differences in chronic diseases characteristics. As assumed, there exists not an one-size-fits-all approach for all types of chronic diseases, but using specific clinical features per disease can deal with the existing different characteristics to a certain extent.

Our future research will drill-down to discover which other chronic diseases can be detected using sleep-wake behavior analysis, and why some of them work less accurate. For different chronic diseases what are the salient features as well as what clinical data play a more significant role in helping disease detection.

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 676157. We thank Dr. Feng Chen for assistance and comments that improved the manuscript.

REFERENCES

[1] H. Tunstall-Pedoe, "Preventing chronic diseases. a vital investment: Who global report. geneva: World health organization," *International Journal of Epidemiology*, vol. 35, no. 4, 2006.

[2] M. E. Falagas, K. Z. Vardakas, and P. I. Vergidis, "Underdiagnosis of common chronic diseases: prevalence and impact on human health," *International Journal of Clinical Practice*, vol. 61, no. 9, pp. 1569–1579, 2007.

[3] A. S. Levey, J. Coresh, E. Balk, A. T. Kausz, A. Levin, M. W. Steffes, R. J. Hogg, R. D. Perrone, J. Lau, and G. Eknoyan, "National Kidney Foundation Practice Guidelines for Chronic Kidney Disease: Evaluation, Classification, and Stratification," *Ann Intern Med.*, vol. 139, pp. 137–147, 2003.

[4] National Heart, Lung and Blood Insitute, U.S. Department of Health and Human Services, "Apnea: Diagnosis-treatment," https://www.nhlbi.nih.gov/health-topics/sleep-apnea, accessed: 2018-04-19.

[5] K. L. Knutson, "Role of Sleep Duration and Quality in the Risk and Severity of Type 2 Diabetes Mellitus," *Archives of Internal Medicine*, vol. 166, no. 16, pp. 1768–74, 2006.

[6] D. Foley, S. Ancoli-Israel, P. Britz, and J. Walsh, "Sleep disturbances and chronic disease in older adults: Results of the 2003 National Sleep Foundation Sleep in America Survey," *Journal of Psychosomatic Research*, vol. 56, no. 5, pp. 497–502, 2004.

[7] C. Thomas-Hawkins and D. Zazworsky, "Self-Management of Chronic Kidney Disease," *American Journal of Nursing*, vol. 105, no. 10, pp. 40–48, 2005.

[8] D. A. Calhoun and S. M. Harding, "Sleep and Hypertension," *Chest - American College of Chest Physicians*, vol. 138, no. 2, pp. 434–443, 2010.

[9] L. Matuzaki, R. Santos-Silva, E. C. Marqueze, C. R. de Castro Moreno, S. Tufik, and L. Bittencourt, "Temporal sleep patterns in adults using actigraph," *Sleep Science*, vol. 7, no. 3, pp. 152–157, 2014.

[10] T. J. Lyons and A. Basu, "Biomarkers in diabetes: hemoglobin a1c, vascular and tissue markers," *Translational Research*, vol. 159, no. 4, pp. 303–312, 2012.

[11] Mayo Clinic, "High blood pressure (hypertension): Diagnosis," https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/diagnosis-treatment/drc-20373417, accessed: 2018-04-19.

[12] C. Laske, H. R. Sohrabi, S. M. Frost, K. Lpez-de Ipia, P. Garrard, M. Buscema, J. Dauwels, S. R. Soekadar, S. Mueller, C. Linnemann, S. A. Bridenbaugh, Y. Kanagasingam, R. N. Martins, and S. E. O'Bryant, "Innovative diagnostic tools for early detection of alzheimer's disease," *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, vol. 11, no. 5, pp. 561–578, 2015.

[13] Y.-E. S. Ju, J. S. McLeland, C. D. Toedebusch, C. Xiong, A. M. F. Fagan, S. P. Duntley, J. C. Morris, and D. M. Holtzman, "Sleep quality and preclinical alzheimer disease," *JAMA Neurology*, vol. 70, no. 5, pp. 587–593, 2013.

[14] K. Aggarwal, S. R. Joty, L. Fernández-Luque, and J. Srivastava, "Comorbidity exploration on wearables activity data using unsupervised pre-training and multi-task learning," *Computing Research Repository*, 2017.

[15] P. D. Sorlie, L. M. Avilés-Santa, S. Wassertheil-Smoller, R. C. Kaplan, M. L. Daviglus, A. L. Giachello, N. Schneiderman, L. Raij, G. Talavera, M. Allison, L. LaVange, L. E. Chambless, and G. Heiss, "Design and implementation of the hispanic community health study/study of latinos," *Ann Epidemiol.*, vol. 20, no. 8, pp. 629–641, 2010.

[16] M. Louter, J. B. Arends, B. R. Bloem, and S. Overeem, "Actigraphy as a diagnostic aid for rem sleep behavior disorder in parkinson's disease," *BMC Neurology*, vol. 14, no. 1, 2014.

[17] Z. C. Lipton, D. C. Kale, C. Elkan, and R. C. Wetzel, "Learning to diagnose with LSTM recurrent neural networks," *Computing Research Repository*, 2015.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computations*, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] F. A. Gers, J. A. Schmidhuber, and F. A. Cummins, "Learning to forget: Continual prediction with lstm," *Neural Computions*, vol. 12, no. 10, pp. 2451–2471, 2000.

[20] F. A. Gers and J. Schmidhuber, "Recurrent nets that time and count," in *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, vol. 3, 2000, pp. 189–194.

[21] A. Graves, "Generating sequences with recurrent neural networks," *Computing Research Repository*, 2013.

[22] Google Code, "Word2vec," https://code.google.com/archive/p/word2vec/, accessed: 2018-08-23.

[23] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation." *CoRR*, 2013.

[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization." *CoRR*, 2014.

[26] National Sleep Research Resource, "Free research data and tools." https://sleepdata.org, accessed: 2018-05-10.

[27] S. M. Grundy, "Pre-diabetes, metabolic syndrome, and cardiovascular risk," *Journal of the American College of Cardiology*, vol. 59, no. 7, pp. 635–643, 2012.

[28] Diabetes.co.uk - the global diabetes community, "Diabetes and sleep," https://www.diabetes.co.uk/diabetes-and-sleep.html, accessed: 2018-05-02.

[29] S. B. Ghaderian and S. S. Beladi-Mousavi, "The role of diabetes mellitus and hypertension in chronic kidney disease," *J Renal Inj Prev.*, vol. 3, no. 4, pp. 109–110, 2014.

[30] J. A. Pinto, D. K. Ribeiro, A. F. d. S. Cavallini, and G. S. Duarte, Caue andFreitas, "Comorbidities associated with obstructive sleep apnea: a retrospective study," *International Archives of Otorhinolaryngology*, vol. 20, no. 2, pp. 145–150, 2016.

[31] M. Gatineau, C. Hancock, N. Holman, H. Outhwaite, L. Oldridge, A. Christie, and L. Ells, "Public health england adult obesity and type 2 diabetes," https://www.gov.uk/government/publications/adult-obesity-and-type-2-diabetes, 2014, accessed: 2018-05-01.

[32] M.-J. Kim, N.-K. Lim, S.-J. Choi, and H.-Y. Park, "Hypertension is an independent risk factor for type 2 diabetes: the korean genome and epidemiology study," *Hypertension Research*, vol. 38, no. 11, pp. 783–789, 2015.

[33] M. Deo, P. Pawar, S. Kanetkar, and S. Kakade, "Prevalence and risk factors of hypertension and diabetes in the Katkari tribe of coastal Maharashtra," *Journal of Postgraduate Medicine*, vol. 63, no. 2, pp. 106–113, 2017.

[34] L. C. Gallo, F. J. Penedo, M. Carnethon, C. Isasi, D. Sotres-Alvarez, V. L. Malcarne, S. C. Roesch, M. E. Youngblood, M. L. Daviglus, P. Gonzalez, and G. P. Talavera, "The hispanic community health study/study of latinos sociocultural ancillary study: Sample, design, and procedures," *Ethnicity & disease*, vol. 24, no. 1, pp. 77–83, 2014.