Towards Autonomous Physiological Signal Extraction From Thermal Videos Using Deep Learning

Kapotaksha Das
takposha@umich.edu
Computer and Information Science
University of Michigan-Dearborn
USA

Mohamed Abouelenien
zmohamed@umich.edu
Computer and Information Science
University of Michigan-Dearborn
USA

Mihai Burzo
mburzo@umich.edu
Mechanical Engineering
University of Michigan-Flint
USA

John Elson
jelson3@ford.com
Ford Motor Company
USA

Kwaku Prakah-Asante
kprakaha@ford.com
Ford Motor Company
USA

Clay Maranville
cmaranvi@ford.com
Ford Motor Company
USA

ABSTRACT

Using the thermal modality in order to extract physiological signals as a noncontact means of remote monitoring is gaining traction in applications, such as healthcare monitoring. However, existing methods rely heavily on traditional tracking and mostly unsupervised signal processing methods, which could be affected significantly by noise and subjects' movements. Using a novel deep learning architecture based on convolutional long short-term memory networks on a diverse dataset of 36 subjects, we present a personalized approach to extract multimodal signals, including the heart rate, respiration rate, and body temperature from thermal videos. We perform multimodal signal extraction for subjects in states of both active speaking and silence, requiring no parameter tuning in an end-to-end deep learning approach with automatic feature extraction. We experiment with different data sampling methods for training our deep learning models, as well as different network designs. Our results indicate the effectiveness and improved efficiency of the proposed models reaching more than 90% accuracy based on the availability of proper training data for each subject.

KEYWORDS

multimodal dataset, deep learning, machine learning, physiological signals, thermal imaging, realtime prediction

ACM Reference Format:
Kapotaksha Das, Mohamed Abouelenien, Mihai Burzo, John Elson, Kwaku Prakah-Asante, and Clay Maranville. 2023. Towards Autonomous Physiological Signal Extraction From Thermal Videos Using Deep Learning. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 09–13, 2023, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3577190.3614123

ICMI '23, October 09–13, 2023, Paris, France

ACM ISBN 979-8-4007-0055-2/23/10…$15.00
https://doi.org/10.1145/3577190.3614123

1 INTRODUCTION

There is a growing interest in developing methods that can effectively and reliably monitor human state. This is usually realized by collecting physiological and behavioral measurements from individuals. For instance, the ability to identify a driver's circadian state in a vehicle using physiological indicators is considered critical to improving the well-being of the vehicle's occupants as well as the quality of the road trip, a growing relevance as we pivot to autonomous vehicles. In addition to circadian state analysis [8], other applications benefiting from such systems include healthcare monitoring in infants [44], detection of emotion [12], detection of fatigue [28], among others.

However, a major challenge in developing such systems is concerned with using contact-based physiological sensors which are inconvenient and can cause discomfort and significant distraction when applied in vehicles [7, 26]. Alternative, non-contact methods started using thermal imaging to demonstrate that there is a high correlation between the non-contact physiological signals found in the thermal data frames, such as the heart rate and respiration rate, and the contact-based ground truth standard measurements [18]. This highlights the utility of the thermal modality as a singular non-contact modality that could be used to monitor multiple aspects of a subject's physiological state, which in turn allows for usage in multiple applications, such as in the detection of cognitive distraction [31] and stress [42].

The flurry of interest in using thermal imaging as a non-contact modality for monitoring can be attributed to its key benefits, the first being that of a robust detector of body temperature [1]. Secondly, it is much more resistant to changes in lighting conditions and will operate even in smoky and dusty environments [17, 30], which is not possible using regular visual systems. These benefits allow for thermal images to contain more complementary information compared to data provided by RGB images, giving it the potential to be used for a variety of applications where the visual modality alone does not suffice. Thirdly, the thermal modality contains information about multiple distinct physiological signals, such as heart rate, respiration rate, and body temperature, which allows for the extraction of multimodal output that would otherwise traditionally require multiple contact-based sensors.

However, there are also challenges utilizing the thermal modality itself and, in particular, the lack of thermal-based datasets and the

inapplicability of existing pre-trained RGB-based models on thermal images due to the nature of the thermal images given that each pixel represent a specific temperature. In RGB image data, there are usually three channels of color-based information per pixel of an image, with color changes often indicating basic features such as boundaries to be established, consequently allowing for more complex features such as shapes to be formed. This does not apply to thermal image data as a thermal image is similar to a 3-D continuous curve, with no clear boundaries between features on the face, for example. This also means that the process of extracting useful information from a thermal frame using deep learning models varies significantly from that of an RGB image.

Furthermore, current unsupervised and signal filtering approaches, which are the methods mainly used to extract physiological signals from thermal images [14, 20, 22, 43], are often dependent on multiple stages of preprocessing and cleaning as well as human parameter selection and tuning, and this makes them susceptible to noise and errors in data when used in scenarios outside the expected operating range [10, 29]. Motivated by the aforementioned challenges and by using the thermal pixel representation along with the increasing computational capacities of deep learning networks, we can leverage the thermal modality in end-to-end applications that do not require any preprocessing or human intervention, which allows for implementations of lightweight monitoring setups that run off cost-effective hardware [13] in a variety of situations. Furthermore, these systems, once trained, provide efficient predictions where costs, space, and discomfort of using contact-based sensors, among other factors, would have caused limited deployment.

Furthermore, there is a growing interest in building personalized models that are more reliable in monitoring subjects as well [9, 21]. Issues such as hidden biases in datasets and reduced sensitivity to individual needs are challenges that exist in big data [3]. Hence, these challenges provide motivation towards developing models that can use user-specific data towards improving predictions made for the said user.

In this paper, we present a novel user-dependent approach toward noncontact physiological signal extraction that does not require Region of Interest (ROI) tracking or signal processing in an end-to-end deep learning pipeline. This approach relies on having separate training data available from the same test subjects to automatically extract relevant features in order to develop personalized models. In particular, we make the following contributions:

(1) A dataset of thermal and physiological recordings of 36 subjects (24 male, 12 female) consisting of over nine hours of data,
(2) A deep learning approach that generates multimodal physiological signal output from a unimodal thermal input,
(3) The implementation of convolutional long short-term memory networks to design a novel architecture that processes the thermal input data directly without parameter tuning,
(4) Generation of heart rate, respiration rate, and body temperature using short windows for efficient predictions.

2 RELATED WORK

There has been a lot of research on the usage of physiological sensors to monitor human state. In a review of wearable sensor technology by Mukhopadhyay [33], the use of body temperature and heart rate are noted as key physiological features to monitor and detect stress, heart attack, and stroke risk, and activity detection. Nicolò et al. [36] in their review of respiration as a physiological signal identified a wide variety of monitoring goals, such as sleep apnea, post-operative patient monitoring, and exercise monitoring, among others. A variety of applications in patient care as well as everyday comfort-of-life scenarios have been established using the ability to monitor physiological signals [5, 32].

The usage of thermal imaging exclusively for physiological signal extraction is a relatively new direction for a non-contact based approach, especially in the context of deep learning. Work by Barbosa et al. [6] noted that in medical settings, attaching physical sensors often resulted in discomfort and stress to the subjects involved, while also noting that there was a high correlation in heart rate and respiration data extracted using thermal imaging methods and gold standard data. Hall et al. [15] also noted that wearable sensors suffered from having finite battery life and stated that contact-based sensors could not be easily used with newborns and burn victims. Thermal imaging has been very useful in detecting human health matters in a non-contact setting [25]. It needs no radiation source for monitoring, unlike in the example of the visual modality needing external lighting, allowing for monitoring in low-light scenarios as well.

A common physiological signal extracted from the thermal modality is the respiration rate. Researchers have explored the usage of thermal cameras to monitor respiration in a sleeping subject non-obtrusively [2, 4, 20, 43], showing that thermal imaging methods could reliably perform better than visual counterparts in situations where lighting and visibility were compromised, while also providing a better monitoring experience for the subject. Elphick et al. [14] demonstrated that while the thermal modality alone was viable for respiration signal extraction, they were still required to manually adjust parameters for better predictions.

Detecting heart rate as a physiological signal has also been of interest, based on the assumption of correlations existing between the thermal data and the pulse rate. However, limited research exists in this area. Lewandowska et al. [27] presented a potential application and algorithm for detecting the heart rate of a motionless subject by using a combination of visual and thermal video recordings using ROIs across the face and Principal Component Analysis. Similarly, Kim et al. used thermal imaging for heart rate detection using temperature changes picked up from blood flow in recorded thermal images [22]. Furthermore, Perpetuini et al. [38], in their research towards the viability of using thermal imaging for heart rate estimation, observed that the support vector machines classifier was capable of estimations that had a high correlation with the ground truth. Nakayama et al. [34] leveraged the use of a CMOS camera-equipped infrared camera system to conduct real-time thermal and visual image processing to measure heart rates and respiration rates that could allow for more sensitive infection detection at airports. To remove the effects of lighting conditions, Hu et al. [19] proposed a dual infrared-thermal camera system that could record subjects sleeping and determine their breathing and heart rates. This demonstrated the robustness of thermal imaging

in various lighting conditions. However, their work did not provide real-time monitoring and similar to other approaches, needed tuning of preprocessing parameters.

There have been some deficiencies in the approaches that have been used in the past for noncontact physiological signal extraction, namely being that the proposed extraction techniques relied on some form of signal preprocessing, which would restrict their ability to reliably operate on a wider scale without needing further tuning and adjustments. Changes in subject behavior or movement could disrupt the effectiveness of these approaches as well. Moreover, there is not as much research into heart rate signal extraction compared to respiration when using the thermal modality, owing to additional challenges, with algorithms being adversely affected by movement, monitoring time and availability of reliable features to track [6]. To counter the problem of feature tracking, some approaches used ROI tracking to monitor regions such as the nostrils or regions with high blood flow [22, 27]. However, these tracking algorithms add computational overhead and might be significantly affected by the tracking performance. Hence, there has been a need and incentive to leverage deep learning approaches that can extract useful features independently and allow for generation of multiple physiological signals from the same thermal data.

Only very recently, a deep learning approach was suggested for this task by Shu et al. [41] where they used models to track the nostrils. Kwasniewska et al. [24] noted how deep learning models extracted features from data differently compared to traditional methods that involved color magnification and enhancements to the data, with better potential using low-resolution data as well. Navneeth et al. [35] used a convoluted neural network (CNN) architecture to detect respiration rate. Similarly, Chen et al. [11] demonstrated that CNN models could be used in detecting heart rate. These approaches do not fully realize the potential of deep learning models as they are implemented only as a part of the signal extraction pipeline that includes ROI tracking or has a limited classification scope. Hence, in our research, we present a deep learning pipeline that can extract multiple physiological signals from the same thermal video using an end-to-end deep learning pipeline.

3 DATASET COLLECTION AND DESCRIPTION

We collected data from 36 consenting subjects, consisting of 24 male and 12 female participants between the ages of 18 to 32 of varying demographics. Data were collected across multiple modalities, captured across three sessions per subject in a climate-controlled lab room to maintain consistency. During the first session, the subject sat silently with natural breathing for two minutes. Following the first session, the subject then participated on another day in two sessions, one in the morning and another in the evening. Each recording session consisted of two parts, a five-minute segment where the subject was told to speak freely about any topic of their choosing, and a two-minute segment where they remained silent, similar to the first session. This allowed us to record subjects during different days and different times in states of activity as they spoke and while they were inactive. For our research, we utilized four channels in particular from the dataset:

(1) Thermal Modality - A FLIR SC7600 thermal camera, capable of recording the subject's face at 100 fps at a resolution of 640x512 pixels representing temperature values were used to capture thermal videos. This camera was situated approximately 42 inches away from the subject, angled upwards at an angle of 18°. This provided us with the thermal video frames that we use as the input to our models.
(2) Physiological Modality - A Biograph Infiniti physiological modality recording suite by Thought Technology was used and utilized multiple sensors to capture three physiological signals:
    (a) Blood Volume Pulse (BVP) - Heart rate / minute - 2048Hz,
    (b) Respiration Rate - No. of breaths / minute - 256 Hz upscaled to 2048Hz, and
    (c) Skin Temperature - Fahrenheit (°F) - 256 Hz upscaled to 2048Hz.

    The respiration rate sensor was attached to the subject's torso and the other sensors were attached to the subject's non-dominant hand. These sensors resulted in three signals of physiological data to provide the ground truth for our models.

These recordings provided approximately nine hours' worth of data to work with. Tables 1 and **??** outline the mean signals for the three physiological signals that we collected.

4 PRE-TRAINING PIPELINE SETUP

Due to the high resolution and frame rate of the thermal images captured, there is a large data bandwidth we had to account for while setting up the pipeline for deep learning. Optimizations and scheduling were hence important parts of this pipeline. Our first task was to convert the collected data from the thermal and physiological modalities to a data format suitable for TensorFlow to set up our training, testing, and validation datasets and their respective labels. More details about the dataset split are provided in Section 5.

As discussed in Section 3, we collected the physiological data across three signals. We synchronized them with the thermal video frames while taking into account two key issues. The first is the different frame rates for each modality. The Bioinfiniti software automatically upscales the sampling rate of the respiration and skin temperature sensors to match that of the heart rate sensor which is beyond our control. The second is the high thermal frame rate that causes the size of a thermal video of just a few seconds to be too large to allow for feasible training and testing. For example, eight seconds of raw thermal video results in over 500 MB of data in size. Hence, we downscaled the frame rate to 8 fps and 8 Hz for all modalities, a sampling rate that reduced the data size and resemble an implementable scenario in vehicles and other applications, where there would be limited hardware, storage, and computational capabilities.

Next, we split the synchronized data into 8-second segments. We chose eight seconds as our segment length as it was the most feasible setting to use in our deep learning pipeline in terms of data size and the available information encoded within it. In our case, at 8 fps, there would be 64 frames per segment. We applied a 50% skip to the starting frame for each segment, allowing us to generate twice as many samples for training had we only used sequential segments of frames that had no skips present. Here, a

Table 1: Statistics for Physiological Signals Across Different Recording Sessions

	Heart Rate (bpm)	Respiration (breaths/min)	Temperature (F)
Baseline	82.8	15.48	88.48
Morning	87.2	14.11	86.74
Evening	84.8	14.08	87.88

Table 2: Statistics for Physiological Signals Across Different Recording Types

	Heart Rate (bpm)	Respiration (breaths/min)	Temperature (F)
Baseline	82.8	15.48	88.48
Active	86.92	13.52	87.48
Silent	83.76	15.78	86.67

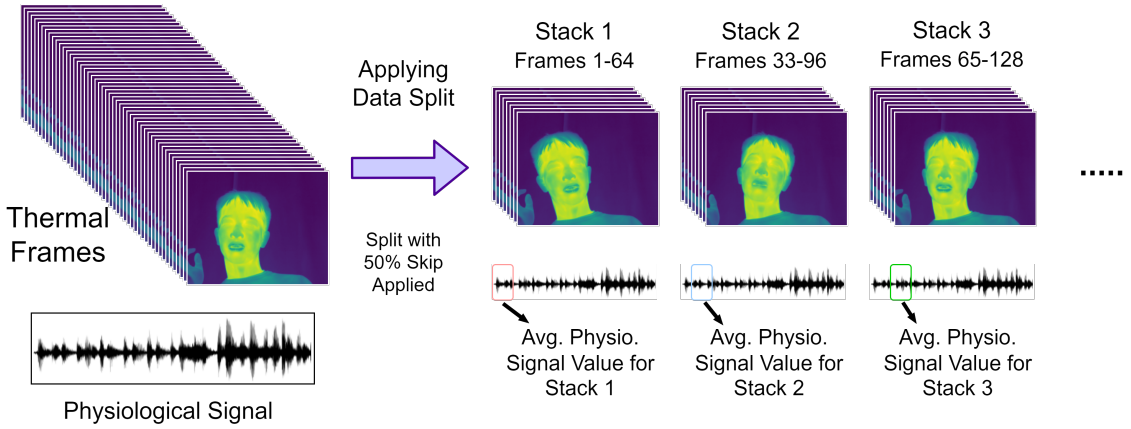


Figure 1: Example of splitting recordings into 8-second stacks

50% skip means that the next segment will start from 50%, or half of the preceding segment. In other words, it starts after the first four seconds of the preceding 8-second segment. An example of the data split can be seen in Fig. 1. We obtained 8393 segments through this method. Some segments had to be discarded if the corresponding ground truth physiological data was corrupted or outside the normal human range due to the subject's movement during the recording. Hence, we finally had a total of 8138 segments that we could use.

While creating the segments, we applied a threshold to the temperature values in the frames, where pixels present in the lower 40% percentile of all pixels values in the frame were set to zero. This allowed us to eliminate large parts of the background to a single value of 0, leaving most pixels on the subject's face unaffected, as they represented higher temperatures. Accordingly, there were no gradients or temperature changes outside the subject's face, which would hypothetically make it easier for a model to focus more on the temperature gradients on the face rather than trying to find and establish an intuition about the background not being useful for predictions. Furthermore, applying the thresholding operation is not computationally expensive, and allows for efficient test predictions to be made.

For each segment, we calculated the mean and variance of the thermal values seen. This was used in order to normalize the data from the frames. We also calculated the mean value of the physiological ground truth, one for each of the three signals per segment. These segment mean values for the physiological data were then scaled to be between 0 and 1, which is preferred when using a sigmoid activation function for the prediction nodes in our models [16]. Hence, the scaled physiological mean value for each segment was used as the label that our models aimed at predicting. No information about the subject, the recording, or the segment's position in a recording was added to the final data in order to remove any information bias that could occur as a result. At the end of this pipeline, we obtained 8138 8-second segments in total to use for training, validation, and testing. Each segment had 64 thermal frames of 512x640 resolution with each pixel representing a temperature value to form a matrix shape of (64,512,640,1).

Using a batch size of four randomly selected thermal video segments, the input stack shape would be (4,64,512,640,1). Accordingly, the four videos could be from different subjects and different recordings. We used a batch size of four to use the maximum available memory on the system used for training, whose specifications are detailed in Section 5. Batching the data was only used in training to optimize the process. In testing, the input data was only one thermal segment at a time, similar to how data would be received in a real-time stream. Fig. 2 gives a representation of the dimension of the input data being supplied to the models.

5 EXPERIMENTAL SETUP

In our experiments, we used two setups. This first is a 'record-split' setup, where different segments from each recording were used in the training and testing sets. In particular, we applied a train-validation-test on our data in the ratio of 60-20-20, sampled randomly from our segments. The second is a 'session-split' setup, where the first session and one of the two other sessions selected randomly of each subject's recordings were used for training and validation, and the remaining session with its two recordings was

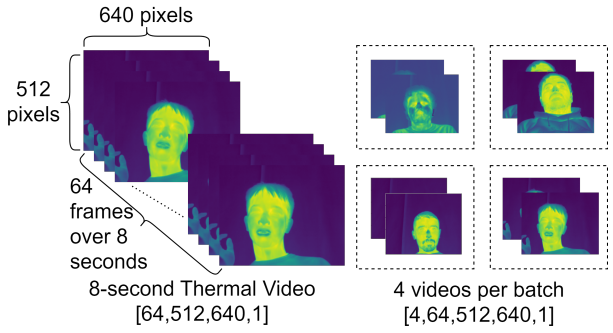


Figure 2: Representation of shape of input data

used for testing, while making sure that the active and silent recordings are included in both the training and testing sets. Using this setup, the train-validation-test split is in the ratio of 45-15-40. Given the duration of the active and silent recordings, the number of samples of the active recordings was much higher. The validation set is used by TensorFlow in order to tune the model parameters, prevent overfitting and allow for the early stopping of model training once the performance has stabilized. We verified that all segments were unique to each set with no overlap, and also ensured that the same data splits were used for each physiological signal being modeled to allow for consistent predictions.

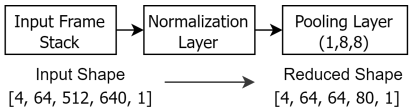
In some preliminary experiments, we tried using TensorFlow models such as ResNet and InceptionV3 which are widely used and prominent visual image processing models on our thermal data. However, these failed to perform well, as it appears that there are certain properties of the thermal data and the physiological signals of interest that these deep learning (DL) models cannot fully extract and represent correctly, as we discussed in Section 1.

We constructed our models using convolutional layers used in convolutional neural networks (CNNs) [37] and long short-term memory (LSTM) networks [39]. Our choice was based on CNNs being networks that are optimal for feature extraction from images using filters that mimic the behavior of biological neurons in the visual cortex to respond to a specific stimulus. These networks require little preprocessing as the filters are optimized during the training phase, drastically reducing the need for human feature engineering. Moreover, as LSTM networks are a form of recurrent neural networks that utilize feedback connections which allows them to process data sequences such as videos, LSTM layers were used to retain information from earlier thermal frames to make better predictions at the current time step. Accordingly, they were used to gain and learn long-term temporal and contextual information, gaining insight into a hierarchical decomposition that might exist in the data.

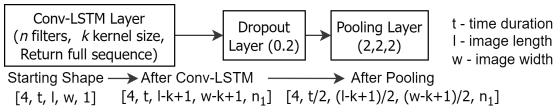
To summarize, we implemented a convolutional LSTM architecture in our models, through the usage of Conv-LSTM layers in Tensorflow [40]. These layers involve the use of convolutional layer structures within the LSTM network itself. We chose this approach over a 3D convolutional (CNN) approach, as a 3D CNN layer might be effective at extracting features in a 3D space, but it loses information about the temporal aspect of the frames. This means we could be losing vital intuition that the thermal values in the frames are continuous across time, corresponding to the movement of the subject's head. Using Conv-LSTM layers allowed for features to be extracted per frame of the stack, while applying a temporal intuition for those features across the stack.

In all of our models, we used a normalization layer to normalize the input frame stack. We performed this using the train mean and variance that we calculated in the pre-training pipeline setup discussed in Section 4. This is good practice as it helps performance during the gradient descent optimization. As using the thermal frames at the original size of 512x640 would result in extremely high memory requirements, we applied a max pooling layer across each frame of the stack, pooling only the spatial dimensions by a factor of 8 as represented in Fig. 3a.

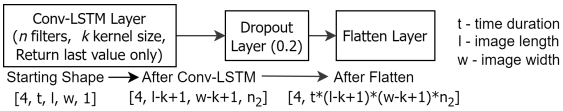
Fig. 3b shows a representation of a standard set of layers that are applied when using Conv-LSTM layers. In models with two or greater Conv-LSTM layers, we returned the full sequence of the stack to be passed to the next layer, as this allowed for layers later in the model to still learn from the temporal aspect of the data.



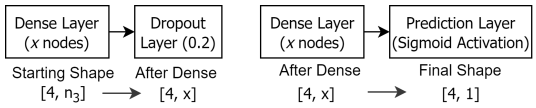
(a) Representation of Input Preprocessing Layers



(b) Representation of Conv-LSTM Layers



(c) Representation of Conv-LSTM Layer with a Flatten Layer



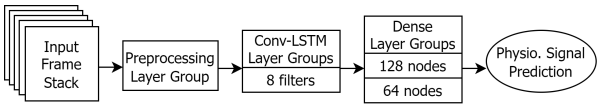
(d) Representation of Dense Layer

Figure 3: Representations of Layers used

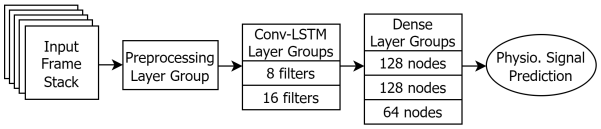
The Conv-LSTM layer before the flatten layer returned only one value instead of a full sequence, as we were interested in modeling the physiological signal for the entire stack as a single value. We applied another dropout layer and then flattened the data so that it can be passed to the dense layers as represented in Fig. 3c.

Finally, we used dense layers as represented in that utilize the features extracted in the Conv-LSTM layers to make the final prediction. For models with two or more dense layers, we also added dropout layers to prevent overfitting. For the very last layer, we used a sigmoid activation to allow for regression, given that we are

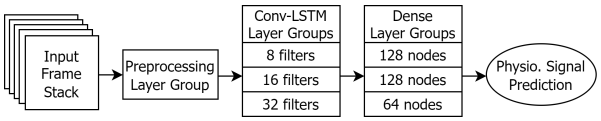
looking for a range of heart and respiration rates as well as skin temperatures.



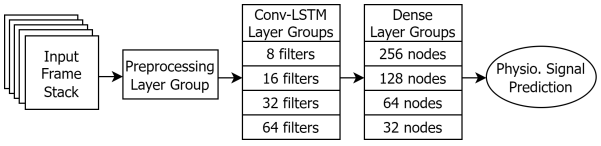
(a) Model Type 1



(b) Model Type 2



(c) Model Type 3



(d) Model Type 4

Figure 4: Representations of Deep Learning Models used

We tested four models by varying the number and sizes of the layer groups discussed earlier to understand how layer complexity can affect the modeling performance for physiological signal extraction. The first as represented in Fig. 4a is the simplest, using only one Conv-LSTM layer followed by two dense layers. The second as represented in Fig. 4b introduces two Conv-LSTM layers and three dense layers. The third model as represented in Fig. 4c uses three Conv-LSTM layers and three dense layers. Finally, the fourth model as represented in Fig. 4d used four Conv-LSTM layers and four dense layers.

We used the mean squared error (MSE) as our metric for estimating the performance of our models on the test split of the data. All training was done using Tensorflow 2.8 on an Intel Xeon 2255-W CPU and Nvidia RTX 3080 GPU, with average training times taking around four hours per epoch. We used the Adam optimizer [23] with a learning rate of 0.001 and implemented an early stopping patience threshold of three epochs to prevent overfitting.

6 EXPERIMENTAL RESULTS AND DISCUSSION

Before diving into the MSE and further analysis of our approach, an important aspect for evaluating performance is the efficiency of our models in terms of the time it takes to process the thermal data and give a prediction. Using our configuration, it took an average of 1.2 seconds to process a test segment and predict a physiological value when given 8 seconds worth of recording. This allows for these models to be applied in near real-time systems, where depending on the amount of computational throughput available, continuous predictions could be made in under 10 seconds of the subject being on frame, or on restrained systems, predictions could be generated in fixed time intervals as well.

Table 3: Mean-Squared Error for Physiological Signals across four Model Types for Record-Split Setup

	Type 1	Type 2	Type 3	Type 4
Heart rate	0.006049	0.005731	**0.005149**	0.006070
Respiration rate	0.002908	**0.001732**	0.002019	0.003765
Temperature	0.002168	0.001665	**0.001536**	0.002873

Table 3 outlines the test MSE observed when using the models to predict the average physiological measurement for a previously unseen 8-second recording segment for the record-split setup. We do not observe significant differences between the performance of the different types of models. For heart rate and skin temperature, the model with three Conv-LSTM layers achieves the lowest error, whereas the respiration rate performs better with two Conv-LSTM layers. It should also be noted that while the MSE scores appear to be quite low and insignificant in the difference between the models, these scores were calculated on a 0 to 1 output range. We discuss more tangible results through error margins in Tables 5 and 6, and Fig. 7 below to give a better understanding of the performance of our models.

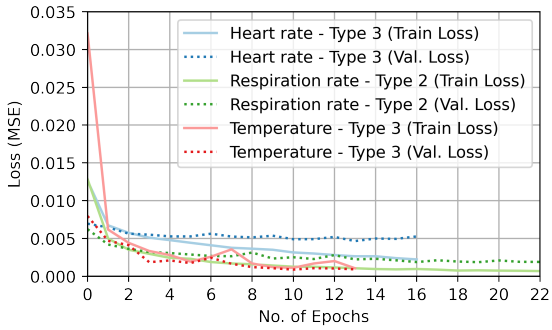


Figure 5: MSE Loss for Best Performing Model per Signal during Training for Record-Split Setup

Fig. 5 shows how each model's loss decreases over the epochs of training for the record-split setup. This gives us an insight into how similarly structured models are trained using very different types of physiological signals. In the case of heart rate, we see that the validation loss converges quickly even as the training loss decreases over time, indicating that despite the model being able to fit the training data better as the number of epochs increases, it does not translate to better predictions on unseen validation data. On the other hand, the model for respiration rate takes the longest number of epochs to train, but both training and validation losses decrease over time, indicating a better capability of the model to fit respiration data and make predictions on unseen data. Lastly, the model for skin temperature has the lowest number of training epochs, and we see that the training loss fluctuates upwards in the second half of the epochs. This is most likely due to the fact

that skin temperature has simpler and non-periodic attributes that correlate easier with thermal temperature data.

Table 4: MSE, R2 and CCC scores for Physiological Signals

Physiological Signal	Record-Split Setup			Session-Split Setup		
	MSE	R2	CCC	MSE	R2	CCC
BVP Heart Rate (Type 3)	0.005	0.765	0.856	0.014	0.341	0.563
Respiration Rate Mean (Type 2)	0.002	0.930	0.962	0.015	0.291	0.555
Temperature (Type 3)	0.002	0.974	0.987	0.036	0.329	0.612

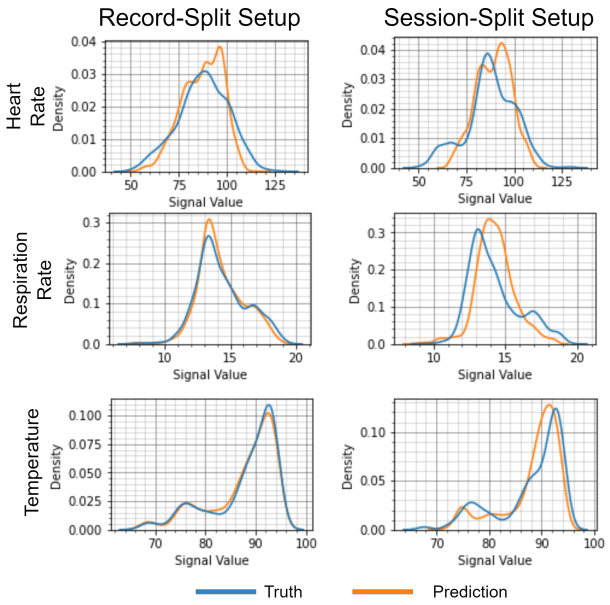


Figure 6: KDE Plots of Ground Truth vs Predicted Signal for Best Performing Model per Signal

MSE scores alone however do not provide a complete picture of the performance. Table 4 tabulates the MSE, Coefficient of determination (R2), and Concordance correlation coefficient (CCC) scores and Fig. 6 plot the KDE distribution of the ground truth versus the predicted signal value for the Record and Session-Split setups for the best-performing model type for each physiological signal. Furthermore, to provide a more tangible understanding of how well our models are performing with respect to accurately predicting physiological signals, we evaluate how close a model's predicted value for a signal is to the ground truth signal value taken from contact-based sensors. Using the best-performing model type for each physiological signal and a 10% margin of error, the results in Table 5 shows that over 87% of all test recordings were predicted correctly within the error margin for heart rate, 98% for respiration rate and 99.7% for temperature for the record-split setup. Table 6 similarly shows the accuracy figures when the data is session-split. As expected, a drop in performance can be seen using this setting given that segments from different days and recording sessions were used for training and testing. However, the results still show a significant improvement over random guessing. In particular, the skin temperature predictions are close to 90%, which was expected given that temperature is primarily encoded in the thermal data.

Table 5: Prediction accuracies for a 10% error margin against real values for Record-Split Setup

	BVP Heart Rate (Type 3)	Respiration Rate Mean (Type 2)	Temperature (Type 3)
Higher than 10% of real value	132	10	3
Within 10% of real value	1421	1597	1624
Lower than 10% of real value	75	21	1
% in range	87.29	98.1	99.75

Table 6: Prediction accuracies for a 10% error margin against real values for Session-Split Setup

	BVP Heart Rate (Type 3)	Respiration Rate Mean (Type 2)	Temperature (Type 3)
Higher than 10% of real value	812	685	294
Within 10% of real value	2082	2145	2932
Lower than 10% of real value	426	490	94
% in range	62.71	64.61	88.31

A comparison of how the models perform over a varying threshold of error margins using both experimental setups can be seen in Fig. 7. In the case of the record-split setup, the models do not stray too far away from the real value within 30% of the real value for the heart rate at worst, or 20% in the case of respiration rate. In the case of a session-split setup, we see that predictions are still within 30 % of the real value for both signals. This shows that the models trained using the session-split setup could potentially benefit from more data segments about the subject in different states. On the other hand, the models trained on randomly sampled data from the same recording have a better intuition of what the expected range of physiological signal values is and consequently make better predictions in that range. This behavior is potentially very useful for allowing models to adapt and personalize to specific users over time, as the ability to adjust to a user's own baseline should allow for better monitoring of deviations in the user as well.

Given that we had subjects recorded in states of both actively speaking and being silent and still, we analyze how our approach performed with regard to the behavior of the subject. Using the same 10% error margin as a way to gauge model performance, we can see the results across the active and silent recording types in Tables 7 and 8. Our models performed significantly better than

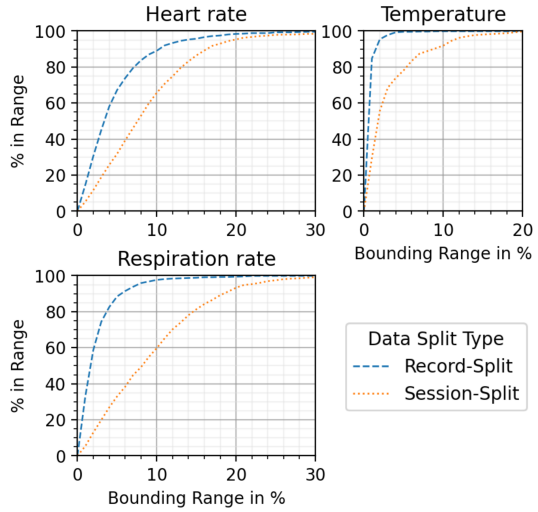


Figure 7: % of Predictions made within Error Margins

Table 7: Prediction accuracies for a 10% error margin for recording split by Active/Silent for Record-Split Setup

Recording Type	Signal	In Range	Low	High	% in Range
Active Recording	Heart Rate	923	67	87	85.7
	Resp. Rate	1053	17	7	99.77
	Temperature	1074	0	3	99.72
Silent Recording	Heart Rate	498	8	45	90.38
	Resp. Rate	544	4	3	98.73
	Temperature	550	1	0	98.2

Table 8: Prediction accuracies for a 10% error margin for activity split by Active/Silent for Session-Split Setup

Recording Type	Signal	In Range	Low	High	% in Range
Active Recording	Heart Rate	1527	392	532	62.3
	Resp. Rate	1609	164	678	65.65
	Temperature	2121	88	242	86.54
Silent Recording	Heart Rate	533	37	299	61.34
	Resp. Rate	516	337	16	59.38
	Temperature	811	6	52	93.33

a baseline model predicting the mean physiological signal value for each segment that was observed during training. The baseline model resulted in accuracy figures of 52% for heart rate, 45% for respiration rate, and 82% for temperature.

In the record-split setup, we observe that the heart and respiration rates are modeled better for silent recordings, despite having a lower number of silent segments during training. This is expected given that our respiration rhythm and heart rate might differ when we speak depending on our speaking rates and the emotions involved. However, this does not hold in the case of the session-split setup. This could be an effect of the recordings being taken at different times, causing individual physiological patterns to also change. Hence, a more effective factor using this setup would be the size of the training data, where we have a larger number of active segments used in training, unlike in the case of the record-split data, where there was more subject-specific information available to better model an individual's physiological characteristics.

7 CONCLUSION

In this paper, we presented a novel user-dependent approach using an end-to-end deep learning approach to extract physiological signals from thermal videos, as a noncontact alternative for monitoring subjects without the need for a contact-based instrument. We used a multimodal dataset of thermal and physiological recordings captured from 36 subjects in states of active speaking and silence to extract heart rate, respiration rate, and body temperature. Our approach relied on having separate training data available from the same subjects used in testing to eliminate the need for ROI tracking or signal processing in order to automatically extract relevant features by implementing a novel personalized architecture using convolutional long short-term memory networks. Using two settings to split the dataset, based on a record-based and session-based data split, our models were able to predict the mean physiological signal value for an 8-second thermal segment in under 1.2 seconds.

We obtained detection accuracy figures of 87% in heart rate, 98% in respiration rate, and 99.7% in body temperature to within 10% of the true signal value recorded when using the record-split setup, and detection accuracy figures of 62.7% in heart rate, 64.6% in respiration rate, and 88.3% in skin temperature to within 10% of the true signal value recorded when using the session-split setup. We observed that respiration is easier to model over heart rate, possibly due to the more complex features needed in extracting useful information for heart rate detection. Furthermore, the type of data split had a significant role in model performance, with the 'record-split' setup performing significantly better with the silent recordings despite having fewer silent segments for training. Conversely, we noted how the larger amount of active data improves performance in the case of a session-split setup for signal predictions. Our results indicate the potential of developing and implementing such personalized models for different applications in vehicles, homes, and healthcare environments that could be reliable in detecting the physiological status of each individual separately.

Our future work would include testing models under other data split setups, such as having different subjects in the training and testing sets, as well as exploring more personalized models. Such models would leverage data recorded across a wider range of times and subject activities to allow for a more precise estimation of different physiological signals remotely.

ACKNOWLEDGMENTS

This material is based in part upon work supported by the Ford Motor Company. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Ford Motor Company or any other Ford entity.

REFERENCES

[1] [n. d.]. Thermal Imaging for Detecting Elevated Body Temperature. https://www.flir.com/discover/public-safety/thermal-imaging-for-detecting-elevated-body-temperature/. Accessed: 2022-03-22.

[2] Abbas K Abbas, Konrad Heimann, Katrin Jergus, Thorsten Orlikowsky, and Steffen Leonhardt. 2011. Neonatal non-contact respiratory monitoring based on real-time infrared thermography. *Biomedical engineering online* 10, 1 (2011), 1–17.

[3] Farhad Ahamed and Farnaz Farid. 2018. Applying Internet of Things and Machine-Learning for Personalized Healthcare: Issues and Challenges. In *2018 International Conference on Machine Learning and Data Engineering (iCMLDE)*. 19–21. https://doi.org/10.1109/iCMLDE.2018.00014

[4] Farah Al-Kalidi, Heather Elphick, Reza Saatchi, and Derek Burke. 2015. Respiratory rate measurement in children using a thermal camera. *International Journal of Scientific and Engineering Research* 6, 4 (2015), 1748–1756.

[5] Andrea Aliverti. 2017. Wearable technology: role in respiratory health and disease. *Breathe* 13, 2 (2017), e27–e36. arXiv:https://breathe.ersjournals.com/content/13/2/e27.full.pdf https://breathe.ersjournals.com/content/13/2/e27

[6] Carina Barbosa Pereira, Michael Czaplik, Vladimir Blazek, Steffen Leonhardt, and Daniel Teichmann. 2018. Monitoring of Cardiorespiratory Signals Using Thermal Imaging: A Pilot Study on Healthy Human Subjects. *Sensors (Basel, Switzerland)* 18, 5 (13 May 2018), 1541. https://doi.org/10.3390/s18051541 29757248[pmid].

[7] Shahina Begum. 2013. Intelligent driver monitoring systems based on physiological sensor signals: A review. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*. 282–289. https://doi.org/10.1109/ITSC.2013.6728246

[8] Clark Bowman, Yitong Huang, Olivia J. Walch, Yu Fang, Elena Frank, Jonathan Tyler, Caleb Mayer, Christopher Stockbridge, Cathy Goldstein, Srijan Sen, and Daniel B. Forger. 2021. A method for characterizing daily physiology from widely used wearables. *Cell Reports Methods* 1, 4 (2021), 100058. https://doi.org/10.1016/j.crmeth.2021.100058

[9] Eli M Cahan, Tina Hernandez-Boussard, Sonoo Thadaney-Israni, and Daniel L Rubin. 2019. Putting the data before the algorithm in big data addressing personalized healthcare. *NPJ Digital Medicine* 2, 1 (2019), 1–6.

[10] Daniela Cardone, Paola Pinti, and Arcangelo Merla. 2015. Thermal infrared imaging-based computational psychophysiology for psychometrics. *Computational and mathematical methods in medicine* 2015 (2015).

[11] Duan-Yu Chen, Huei-Siang Zou, and An-Ting Hsieh. 2020. Thermal Image based Remote Heart Rate Measurement on Dynamic Subjects Using Deep Learning. In *2020 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*. 1–2. https://doi.org/10.1109/ICCE-Taiwan49838.2020.9258129

[12] Irving A. Cruz-Albarran, Juan P. Benitez-Rangel, Roque A. Osornio-Rios, and Luis A. Morales-Hernandez. 2017. Human emotions detection based on a smart-thermal system of thermographic images. *Infrared Physics & Technology* 81 (2017), 250–261. https://doi.org/10.1016/j.infrared.2017.01.002

[13] Robert David, Jared Duke, Advait Jain, Vijay Janapa Reddi, Nat Jeffries, Jian Li, Nick Kreeger, Ian Nappier, Meghna Natraj, Shlomi Regev, Rocky Rhodes, Tiezhen Wang, and Pete Warden. 2020. TensorFlow Lite Micro: Embedded Machine Learning on TinyML Systems. https://doi.org/10.48550/ARXIV.2010.08678

[14] H. E. Elphick, A. H. Alkali, R. K. Kingshott, D. Burke, and R. Saatchi. 2019. Exploratory Study to Evaluate Respiratory Rate Using a Thermal Imaging Camera. *Respiration* 97, 3 (2019), 205–212. https://doi.org/10.1159/000490546

[15] Travis Hall, Donald Y. C. Lie, Tam Q. Nguyen, Jill C. Mayeda, Paul E. Lie, Jerry Lopez, and Ron E. Banister. 2017. Non-Contact Sensor for Long-Term Continuous Vital Signs Monitoring: A Review on Intelligent Phased-Array Doppler Sensor Design. *Sensors* 17, 11 (2017). https://doi.org/10.3390/s17112632

[16] Jun Han and Claudio Moraga. 1995. The influence of the sigmoid function parameters on the speed of backpropagation learning. In *From Natural to Artificial Neural Computation*, José Mira and Francisco Sandoval (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 195–201.

[17] Kirk Havens. 2015. *Thermal Imaging Techniques to Survey and Monitor Animals in the Wild : a Methodology*. Academic Press Imprint,Elsevier Science & Technology Books, San Diego.

[18] Christian Hessler, Mohamed Abouelenien, and Mihai Burzo. 2020. A Non-contact Method for Extracting Heart and Respiration Rates. In *2020 17th Conference on Computer and Robot Vision (CRV)*. 1–8. https://doi.org/10.1109/CRV50864.2020.00009

[19] Menghan Hu, Guangtao Zhai, Duo Li, Yezhao Fan, Huiyu Duan, Wenhan Zhu, and Xiaokang Yang. 2018. Combination of near-infrared and thermal imaging techniques for the remote and simultaneous measurements of breathing and heart rates under sleep situation. *PloS one* 13, 1 (2018), e0190466.

[20] Prasara Jakkaew and Takao Onoye. 2020. Non-Contact Respiration Monitoring and Body Movements Detection for Sleep Using Thermal Imaging. *Sensors (Basel, Switzerland)* 20, 21 (05 Nov 2020), 6307. https://doi.org/10.3390/s20216307 33167556[pmid].

[21] Kevin B. Johnson, Wei-Qi Wei, Dilhan Weeraratne, Mark E. Frisse, Karl Misulis, Kyu Rhee, Juan Zhao, and Jane L. Snowdon. 2021. Precision Medicine, AI, and the Future of Personalized Health Care. *Clinical and Translational Science* 14, 1 (2021), 86–93. https://doi.org/10.1111/cts.12884 arXiv:https://ascpt.onlinelibrary.wiley.com/doi/pdf/10.1111/cts.12884

[22] Yoonkyoung Kim, Yosep Park, and Eui Chul Lee. 2018. Remote Heart Rate Monitoring Method Using Infrared Thermal Camera.

[23] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[24] Alicja Kwasniewska, Jacek Ruminski, and Maciej Szankin. 2019. Improving Accuracy of Contactless Respiratory Rate Estimation by Enhancing Thermal Sequences with Deep Neural Networks. *Applied Sciences* 9, 20 (2019). https://doi.org/10.3390/app9204405

[25] B.B. Lahiri, S. Bagavathiappan, T. Jayakumar, and John Philip. 2012. Medical applications of infrared thermography: A review. *Infrared Physics & Technology* 55, 4 (2012), 221–235. https://doi.org/10.1016/j.infrared.2012.03.007

[26] Steffen Leonhardt, Lennart Leicht, and Daniel Teichmann. 2018. Unobtrusive Vital Sign Monitoring in Automotive Environments—A Review. *Sensors* 18, 9 (2018). https://doi.org/10.3390/s18093080

[27] Magdalena Lewandowska, Jacek Rumiński, Tomasz Kocejko, and Jędrzej Nowak. 2011. Measuring pulse rate with a webcam — A non-contact method for evaluating cardiac activity. In *2011 Federated Conference on Computer Science and Information Systems (FedCSIS)*. 405–410.

[28] Miguel Bordallo Lopez, Carlos R. del Blanco, and Narciso Garcia. 2017. Detecting exercise-induced fatigue using thermal imaging and deep learning. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*. 1–6. https://doi.org/10.1109/IPTA.2017.8310151

[29] Martin Clinton Tosima Manullang, Yuan-Hsiang Lin, Sheng-Jie Lai, and Nai-Kuan Chou. 2021. Implementation of Thermal Camera for Non-Contact Physiological Measurement: A Systematic Review. *Sensors* 21, 23 (2021). https://doi.org/10.3390/s21237777

[30] Jiři Mekyska, Virginia Espinosa-Duró, and Marcos Faundez-Zanuy. 2010. Mekyska. In *44th Annual 2010 IEEE International Carnahan Conference on Security Technology*. 185–189. https://doi.org/10.1109/CCST.2010.5678709

[31] Masahiro Miyaji, Haruki Kawanaka, and Koji Oguri. 2009. Driver's cognitive distraction detection using physiological features by the adaboost. In *2009 12th International IEEE Conference on Intelligent Transportation Systems*. 1–6. https://doi.org/10.1109/ITSC.2009.5309881

[32] Jermana L. Moraes, Matheus X. Rocha, Glauber G. Vasconcelos, José E. Vasconcelos Filho, Victor Hugo C. De Albuquerque, and Auzuir R. Alexandria. 2018. Advances in Photopletysmography Signal Analysis for Biomedical Applications. *Sensors* 18, 6 (2018). https://doi.org/10.3390/s18061894

[33] Subhas Chandra Mukhopadhyay. 2015. Wearable Sensors for Human Activity Monitoring: A Review. *IEEE Sensors Journal* 15, 3 (2015), 1321–1330. https://doi.org/10.1109/JSEN.2014.2370945

[34] Yosuke Nakayama, Guanghao Sun, Shigeto Abe, and Takemi Matsui. 2015. Non-contact measurement of respiratory and heart rates using a CMOS camera-equipped infrared camera for prompt infection screening at airport quarantine stations. In *2015 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*. 1–4. https://doi.org/10.1109/CIVEMSA.2015.7158595

[35] S Navaneeth, S Sarath, B Amba Nair, K Harikrishnan, and P Prajal. 2020. A Deep-Learning Approach to Find Respiratory Syndromes in Infants using Thermal Imaging. In *2020 International Conference on Communication and Signal Processing (ICCSP)*. 0498–0501. https://doi.org/10.1109/ICCSP48568.2020.9182231

[36] Andrea Nicolò, Carlo Massaroni, Emiliano Schena, and Massimo Sacchetti. 2020. The Importance of Respiratory Rate Monitoring: From Healthcare to Sport and Exercise. *Sensors* 20, 21 (2020). https://doi.org/10.3390/s20216396

[37] Keiron O'Shea and Ryan Nash. 2015. An Introduction to Convolutional Neural Networks. *CoRR* abs/1511.08458 (2015). arXiv:1511.08458 http://arxiv.org/abs/1511.08458

[38] David Perpetuini, Andrea Di Credico, Chiara Filippini, Pascal Izzicupo, Daniela Cardone, Piero Chiacchiaretta, Barbara Ghinassi, Angela Di Baldassarre, and Arcangelo Merla. 2021. Is It Possible to Estimate Average Heart Rate from Facial Thermal Imaging? *Engineering Proceedings* 8, 1 (2021). https://doi.org/10.3390/engproc2021008010

[39] Alex Sherstinsky. 2018. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. *CoRR* abs/1808.03314 (2018). arXiv:1808.03314 http://arxiv.org/abs/1808.03314

[40] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1* (Montreal, Canada) *(NIPS'15)*. MIT Press, Cambridge, MA, USA, 802–810.

[41] Shuangbao Shu, Huajun Liang, Yu Zhang, Yuzhong Zhang, and Ziqiang Yang. 2022. Non-contact measurement of human respiration using an infrared thermal camera and the deep learning method. *Measurement Science and Technology* (2022). http://iopscience.iop.org/article/10.1088/1361-6501/ac5ed9

[42] Riccardo Sioni and Luca Chittaro. 2015. Stress Detection Using Physiological Sensors. *Computer* 48, 10 (2015), 26–33. https://doi.org/10.1109/MC.2015.316
[43] Muhammad Usman, Ruth Evans, Reza Saatchi, Ruth Kingshott, and Heather Elphick. 2019. Non-invasive respiration monitoring by thermal imaging to detect sleep apnoea. (2019).
[44] Mauricio Villarroel, Sitthichok Chaichulee, João Jorge, Sara Davis, Gabrielle Green, Carlos Arteta, Andrew Zisserman, Kenny McCormick, Peter Watkinson, and Lionel Tarassenko. 2019. Non-contact physiological monitoring of preterm infants in the Neonatal Intensive Care Unit. *npj Digital Medicine* (2019).