Man Overboard: Fall detection using spatiotemporal convolutional autoencoders in maritime environments

Nikolaos, NB, Bakalos
National Technical University of Athens, Athens, Greece
bakalosnik@mail.ntua.gr

Iason, IK, Katsamenis
National Technical University of Athens, Athens, Greece
iasonkatsamenis@mail.ntua.gr

Athanasios, AV, Voulodimos*
University of West Attica, Athens, Greece
avoulod@uniwa.gr

ABSTRACT

Man overboard incidents in a maritime vessel are serious accidents where, the efficient and rapid detection is crucial in the recovery of the victim. The severity of such accidents, urge the use of intelligent systems that are able to automatically detect a fall and provide relevant alerts. To this end the use of novel deep learning and computer vision algorithms have been tested and proved efficient in problems with similar structure. This paper presents the use of a deep learning framework for automatic detection of man overboard incidents. We investigate the use of simple RGB video streams for extracting specific properties of the scene, such as movement and saliency, and use convolutional spatiotemporal autoencoders to model the normal conditions and identify anomalies. Moreover, in this work we present a dataset that was created to train and test the efficacy of our approach.

CCS CONCEPTS

• **Computing methodologies**; • **Artificial intelligence**; • **Computer vision**; • **Computer vision problems**; • **Object detection**; • **Computer vision tasks**; • **Scene anomaly Detection**;

KEYWORDS

Man overboard, Human detection, Deep learning Computer

ACM Reference Format:
Nikolaos, NB, Bakalos, Iason, IK, Katsamenis, and Athanasios, AV, Voulodimos. 2021. Man Overboard: Fall detection using spatiotemporal convolutional autoencoders in maritime environments. In *The 14th PErvasive Technologies Related to Assistive Environments Conference (PETRA 2021), June 29–July 02, 2021, Corfu, Greece*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3453892.3461326

PETRA 2021, June 29–July 02, 2021, Corfu, Greece

ACM ISBN 978-1-4503-8792-7/21/06…$15.00
https://doi.org/10.1145/3453892.3461326

1 INTRODUCTION

A man overboard is an emergency incident, where a crew member or passenger of a maritime vessel has fallen off-vessel in the sea. These types of accidents are more often in passenger ships, where there is presence of a large number of untrained individuals. It is estimated that 22 people fall off a cruise ship annually [1]. Moreover, these incidents have high mortality rates, as almost 79% of the victims do not survive or are considered missing [1]. The cause of such high mortality rates is the low speed of detection and retrieval. After an hour in water at 4.4{°}C, body temperature drops to 30{°}C [2]. Thus, it is a critical event that demands immediate handling as time plays an important role and because the overboard casualty is exposed to various security risks, such as drowning at sea, hypothermia, injuries and rough sea. It is noted that the problem lies in the lack of timely and critical information, such as the accurate confirmation of the event as well as its exact time and position of the occurrence.

2 PREVIOUS WORK

In a universal maritime surveillance system, human detection is a key issue and must be completely independent of the environment as well as light and weather conditions [3]. Several human detection methods have been presented in the literate and have emphasized the importance of real-time home surveillance systems (e.g. [4], [5]) that focus on fall detection through visual sensors, deep learning and computer vision applications (e.g. [6] - [8]), however, little work has been presented in the literature on the man overboard situation.

In essence, though the incident can be modeled as an abnormal behavior detection problem, where the normal situation consists of a normal capturing of a seafaring vessel, while the abnormality would be the capturing of a fall. To this end, the main approaches for abnormal event recognition involve either the use of supervised deep learning techniques to learn a dictionary of abnormal sub-events or unsupervised outlier detection techniques [9] - [11]. Examples include surveillance in industrial environments [9] or critical infrastructures [11] for safety/security and quality assurance, traffic flow management [12], and intelligent monitoring of public places [13]

Regarding outlier detection, the works of [14] - [16] learn from a dictionary of sub-events, through a training process, and then those events that do not lie in the partitioned sub-space are marked as abnormal ones.

Regarding deep learning, the work of [17] employs convolutional auto-encoders (ConvAE) to learn temporal regularity in videos, while auto-encoders are exploited in [18] to learn features and reconstruct the input images. Then, one-class Support Vector Machines (SVMs) are used for detecting abnormal events. The work

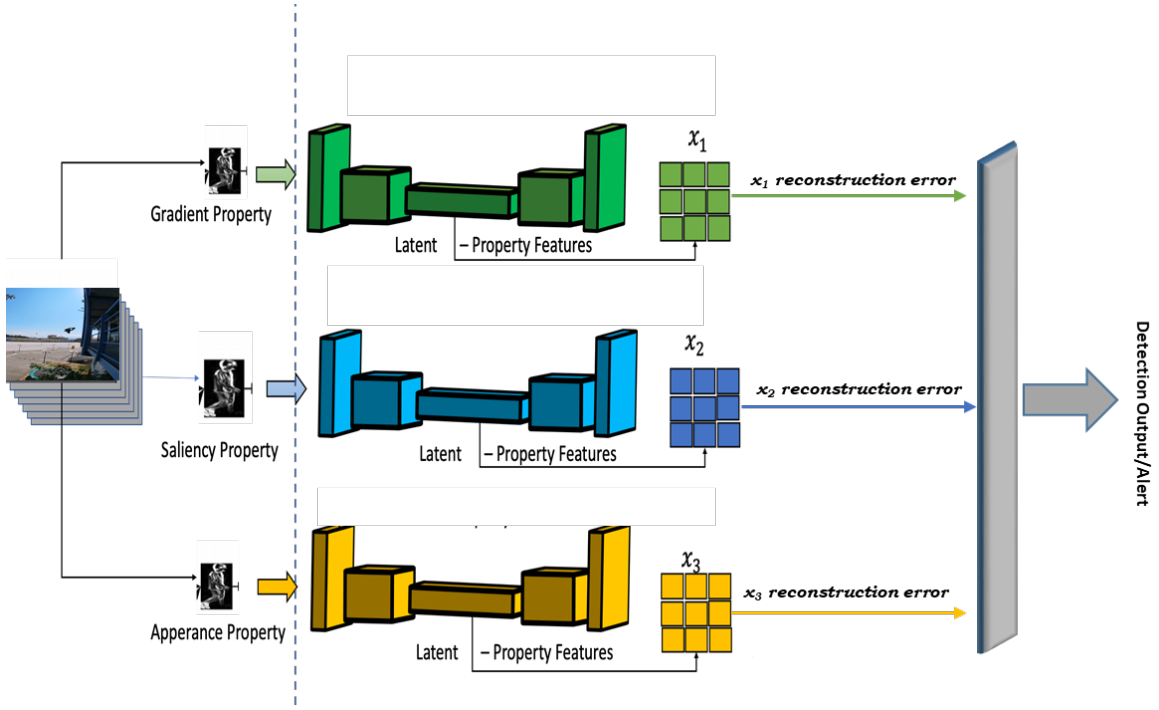

Figure 1: Overall system architecture.

of [19] introduces a hybrid scheme that aggregates ConvAE with Long Short-Term Memory (LSTM) encoder-decoder.

Recently, deep generative models have been applied [17] - [19]. These models are trained to produce normal events while the abnormal ones are given as the difference between the original frames and the generated ones.

In parallel, unsupervised learning models are utilized for abnormal event detection. In [20], the anomalies in videos are scored independently of temporal ordering and without any training by simply discriminating between abnormal frames and normal ones. Other approaches exploit on-line incremental coding [21], deep cascading neural networks, and unmasking (a technique previously used for authorship verification in text documents) [22]. Recently, the works of [23] and [24] incorporate autoencoders and supervised learning for abnormal event detection. Other approaches employ tracking algorithms to extract salient motion information which is then classified either as normal or abnormal [25], [26]. However, tracking fails in complex visual scenes where multiple humans are present.

In this paper, we present the use of an unsupervised fall detection method for man overboard scenarios (see Figure 1). Our approach is based on the use of convolutional spatiotemporal autoencoders trained using a dataset that simulates man overboard incidents. We use multiple image properties as a way to enhance the detection capabilities of our system.

3 SYSTEM ARCHITECTURE

The presented system uses only RGB video streams to identify overboard falls. However, the simple use of raw RGB frames is not sufficient for an efficient detection. To extract additional data from the visual modality we furtherly analyzed the camera streams to extract specific visual properties, i.e. representative vectors. To this end, the visual modality is analyzed furtherly to extract the actual frame (appearance), the gradient of the frame using a short memory window of 10 frames (movement vector), the objectness of the current frame (saliency vector). The Appearance Property consists of the actual frame capturing. The Motion Property captures the movement of objects by taking as input the gradient of the frame. Finally, the Saliency Property reflects how likely a window of the frame covers an object of any category. This property creates a saliency map with the same size as the frame that covers all objects in an image in a category-independent manner.

Each image property was fed into an individual spatiotemporal autoencoder. Autoencoders are a type of Neural Network that manages to learn efficient data encodings by training the network to ignore signal noise. Their usefulness comes from the fact that they are trained in an unsupervised manner. They are essentially composed of two main components that are trained in parallel. The dimensionality reduction component aims at extracting an efficient encoding of the input signal, while the reconstruction side tries to generate from the reduced encoding a representation as close as possible to the original input. To identify the abnormalities, the reconstruction error of each autoencoder was monitored, and when the error was bigger than a predefined threshold, an alert was raised. The selection of the threshold took place during the training, to identify the exact value that maximized detection performance.

The autoencoders used for each image property had the structure presented in Figure 2. Each RGB frame for the appearance vector

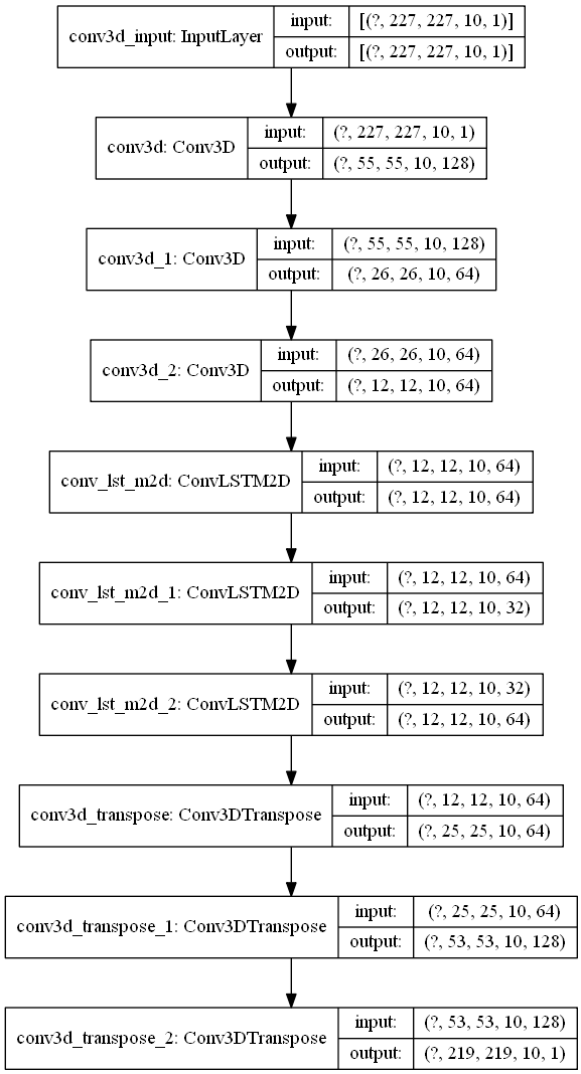

Figure 2: Autoencoder structure.

was reduced to a grayscale image with a resolution of 227x227x1. A 10 frame batch was used for the analysis.

4 PERFORMANCE EVALUATION

4.1 Dataset Description

In order to train and evaluate the proposed methodology, a mock man-overboard event was conducted that concerned the fall of a human-sized dummy from the balcony of a high-rise building. In particular, the human dummy, weighting 30 Kg, was thrown from an approximate height of 20 meters, which is roughly equivalent to two seconds of free-falling. For the needs of the experiment, we made 320 test throws of the dummy, to simulate a man-overboard event (see Figure 3(a)-(d)). Additionally, we recorded several videos without dropping the dummy as well as numerous throws of various objects, such as plastic bags and bottles (see Figure 3(e)). This way we can implement deep learning models that are not prone to false-positive alarms, triggered by non-human-related events.

The experiments took place in the surrounding area of Nikaia Olympic Weightlifting Hall, and lasted five days. Due to the fact that the test throws were carried out throughout the whole day, from 9:00 AM to 5:00 PM, the acquired videos vary in terms of illumination conditions (e.g. underexposure, overexposure). Additionally, we shot under various weather conditions (e.g. sunny, cloudy, rainy, windy, hot, cold), thus providing further variations in the background of the event.

In this paper, we are using a dataset consisted of RGB videos featuring the free falls of the dummy (see Figure 3(a)-(d)). For the dataset collection, which contains video sequences with a resolution of 1080×1920 pixels, we used a GoPro Hero 7 Silver. The camera was set to shoot at a high frame rate, at 50 frames per second, to ensure sufficient acquisition of data that concerns the critical event.

It is underlined that to avoid training bias and guarantee replicability of the results to other datasets, we placed the sensor in four different locations of the building, in order to obtain data that vary in terms of background, illumination, shooting angle, and distance (see Figure 3(a)-(d)). In particular, as depicted in Figure 4, we placed the RGB camera (i) on the left of the fall at a close distance of 7m (see Figure 3(a)), (ii) on the right of the fall at a close distance of 5m (see Figure 3(b)), (iii) on the top left of the fall at an angle of roughly 45{°} (see Figure 3(c)), and (iv) to the left of the fall at a long distance of 13m (see Figure 3(d)). It is emphasized that to further generalize the learning procedure, we augmented the training data by horizontally flipping the corresponding videos.

4.2 Overview of the implementation

The proposed method was implemented in Python. The autoencoders that perform the feature extraction (Appearance, Gradient, and Saliency) were implemented in Tensorflow and Keras. The hyperparameter optimization of the learning algorithms was determined using the Hyperband optimization method of [27], which employs a principled early-stopping strategy to allocate resources, allowing it to evaluate orders-of-magnitude more configurations than black-box procedures like Bayesian optimization methods [28]. The implementation used Python 3.6, Keras (1.08), and Tensorflow (2.1.0) machine learning libraries, in combination with a number of other scientific and data management libraries. The model was trained using an Intel Core i7-6700K CPU (4GHz) with 2 NVIDIA GTX1080 GPUs.

4.3 Experimental Validation

The Area Under Curve (AUC) metric was employed in assessing the performance of the proposed method and the compared ones. The AUC is computed with regard to ground-truth annotations at the frame-level and it is a common metric for many abnormal event detection methods. It measures the ability of the learning algorithm to correctly distinguish normal from abnormal events and summarises the ROC curve of the system, i.e. the probability curve that plots the raising a true alert (true positive rate) and a false alarm (false positive rate) at various thresholds. Our algorithm achieves an AUC score of 97.3. Due to the fact that there are no similar publications for fall detection in man overboard scenarios,

(a) (b) (c) (d) (e)

Figure 3: Test throws during the data collection experiments. The free fall (a)-(d) of the human dummy from different shooting angles (positive event), and (e) of a plastic bag (negative event).

Figure 4: The four locations of the building that the optical sensor was placed, during the data acquisition experiments.

at least to the authors' knowledge, a comparative analysis of the performance is hard to achieve. However, if we considered each frame including a part of a fall as abnormal and all other frames as normal, we can assess the performance of the system using the classification metrics of accuracy, precision, recall and F1-score. The performance of our system using these metrics can be viewed in Figure 5

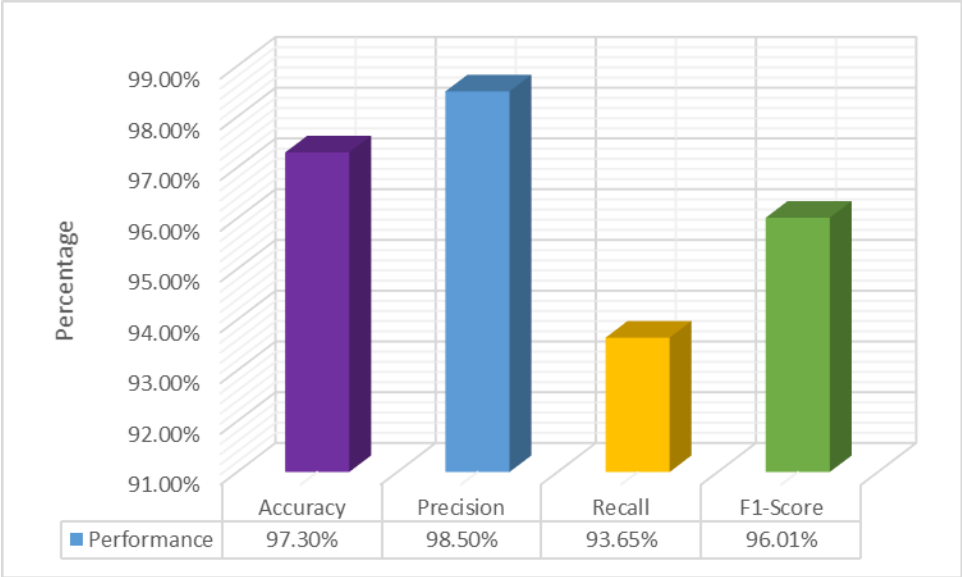

Figure 5: Performance assessment of the proposed learning algorithm.

5 CONCLUSIONS

In this paper, we presented and evaluated a learning algorithm for man overboard detection. The employed techniques use a deep machine learning framework, modeling a man overboard incident as an abnormal action recognition one. The system utilizes multi-property analysis of video streams to extract salient features and encodings of the normal scene using a set of convolutional spatiotemporal autoencoders. We then proceed in identifying falls by the autoencoders' success or failure to reconstruct a scene due to the presence of abnormal events.

Future work should include the presence of additional imaging modules, such as thermal imaging frames, and the studying of additional ways for inter and intra property encoding of all the available modalities to maximize the detection capabilities.

ACKNOWLEDGMENTS

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T1EDK-01169)

REFERENCES

[1] E. Örtlund and M. Larsson, "Man Overboard detecting systems based on wireless technology," 2018.
[2] Sevin, C. BAYILMIŞ, İ. ERTÜRK, H. EKİZ, and A. Karaca, "Design and implementation of a man-overboard emergency discovery system based on wireless sensor networks," Turk. J. Electr. Eng. Comput. Sci., vol. 24, no. 3, pp. 762–773, 2016.
[3] I. Katsamenis , E. Protopapadakis, A. S. Voulodimos, D. Dres, and D. Drakoulis. "Man overboard event detection from rgb and thermal imagery: Possibilities and limitations," In Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments, pp. 1-6. 2020.
[4] A. S. Voulodimos, D. I. Kosmopoulos, N. D. Doulamis, and T. A. Varvarigou, "A top-down event-driven approach for concurrent activity recognition," Multimed. Tools Appl., vol. 69, no. 2, pp. 293–311, Mar. 2014, doi: 10.1007/s11042-012-0993-4.
[5] N. D. Doulamis, A. S. Voulodimos, D. I. Kosmopoulos, and T. A. Varvarigou, "Enhanced Human Behavior Recognition Using HMM and Evaluative Rectification," in Proceedings of the First ACM International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams, New York, NY, USA, 2010, pp. 39–44, doi: 10.1145/1877868.1877880.
[6] K. Makantasis, E. Protopapadakis, A. Doulamis, N. Doulamis, and N. Matsatsinis, "3D measures exploitation for a monocular semi-supervised fall detection system," Multimed. Tools Appl., vol. 75, no. 22, pp. 15017–15049, Nov. 2016, doi: 10.1007/s11042-015-2513-9.
[7] C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau, "Robust video surveillance for fall detection based on human shape deformation," IEEE Trans. Circuits Syst. Video Technol., vol. 21, no. 5, pp. 611–622, 2011.
[8] M. Yu, A. Rhuma, S. M. Naqvi, L. Wang, and J. Chambers, "A posture recognition-based fall detection system for monitoring an elderly person in a smart home environment," IEEE Trans. Inf. Technol. Biomed., vol. 16, no. 6, pp. 1274–1286, 2012.
[9] S. Lee, H. G. Kim and Y. M. Ro, "BMAN: Bidirectional Multi-Scale Aggregation Networks for Abnormal Event Detection," IEEE Trans. on Image Proc., vol. 29, pp. 2395-2408, 2020.
[10] A. S. Voulodimos, N.D. Doulamis, D.I. Kosmopoulos, and T.A. Varvarigou, "Improving multi-camera activity recognition by employing neural network based readjustment," Applied Artificial Intelligence, 26(1-2), 97-118, 2012.
[11] N. Bakalos, *et al.* "Protecting water infrastructure from cyber and physical threats: Using multimodal data fusion and adaptive deep learning to monitor critical systems." IEEE Signal Processing Magazine, 36.2, pp. 36-48, 2019.
[12] S. Wan, X. Xu, T. Wang and Z. Gu, "An Intelligent Video Analysis Method for Abnormal Event Detection in Intelligent Transportation Systems," IEEE Trans. on Intell. Transportation Systems, (to be published)
[13] R. Leyva, V. Sanchez and C. Li, "Fast Detection of Abnormal Events in Videos with Binary Features," IEEE ICASSP, Calgary, AB, pp. 1318-1322, 2018.
[14] K.-W. Cheng, Y.-T. Chen, and W.-H. Fang, "Video anomaly detection and localization using hierarchical feature repre- sentation and Gaussian process regression," IEEE CVPR, pp. 2909–2917, 2015.
[15] C. Lu, J. Shi, and J. Jia, "Abnormal Event Detection at 150 FPS in MATLAB," IEEE ICCV, pages 2720– 2727, 2013.
[16] H. Ren, W. Liu, S. I. Olsen, S. Escalera, and T. B. Moes- lund, "Unsupervised Behavior-Specific Dictionary Learning for Abnormal Event Detection," Proc. of BMVC, pp. 28.1–28.13, 2015.
[17] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," IEEE CVPR, pages 733–742, 2016.
[18] Xu Dan, Ricci Elisa, Yan Yan, Song Jingkuan and Sebe Nicu, "Learning Deep Representations of Appearance and Motion for Anomalous Event Detection", BMVC, 2015.
[19] L. Wang, F. Zhou, Z. Li, W. Zuo and H. Tan, "Abnormal Event Detection in Videos Using Hybrid Spatio-Temporal Autoencoder," 25th IEEE International Conference on Image Processing (ICIP), Athens, 2018, pp. 2276-2280, 2018.
[20] A. Del Giorno, J. Bagnell, and M. Hebert, "A Discrimina- tive Framework for Anomaly Detection in Large Videos," Proc. of ECCV, pp. 334–349, 2016.

[21] J. K. Dutta and B. Banerjee, "Online Detection of Abnormal Events Using Incremental Coding Length," In Proceedings of AAAI, pages 3755–3761, 2015.
[22] R.T. Ionescu, S. Smeureanu, B. Alexe, and M. Popescu, "Un-masking the abnormal events in video," IEEE ICCV, pp. 2895–2903, 2017.
[23] S. Smeureanu, R. T. Ionescu, M. Popescu, and B. Alexe, "Deep Appearance Features for Abnormal Behavior Detection in Video," In Proceedings of ICIAP, Volume 10485, pages 779–789, 2017.
[24] Y. Liu, C.-L. Li, and B. Poczos, "Classifier Two-Sample Test for Video Anomaly Detections," In Proceedings of BMVC, 2018.
[25] X. Mo, V. Monga, R. Bala, and Z. Fan, "Adaptive sparse representations for video anomaly detection," IEEE Trans. Circuits Syst. Video Technol., vol. 24, no. 4, pp. 631–645, Apr. 2014.
[26] F. Jiang, Y. Wu, and A. K. Katsaggelos, "A dynamic hierarchical clustering method for trajectory-based unusual video event detection," IEEE Trans. Image Process., vol. 18, no. 4, pp. 907–913, Apr. 2009.
[27] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A novel bandit-based approach to hyperparameter optimization," The Journal of Machine Learning Research, 18(1), pp.6765-6816, 2016.
[28] Kaselimi, Maria, *et al.* "Bayesian-optimized bidirectional LSTM regression model for non-intrusive load monitoring." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019