Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 38 (2019) 248–255

Procedia MANUFACTURING

www.elsevier.com/locate/procedia

29th International Conference on Flexible Automation and Intelligent Manufacturing (FAIM2019), June 24-28, 2019, Limerick, Ireland.

Deep Learning-based Production Forecasting in Manufacturing: a Packaging Equipment Case Study

Luca Brunelli[a], Chiara Masiero[a], Diego Tosato[b], Alessandro Beghi[c], Gian Antonio Susto[c,*,**]

[a]*Statwolf Data Science SRL, Padova 35131 , Italy,*
[b]*Galdi SRL, Postioma (TV) 31038, Italy*
[c]*University of Padova, Padova 35122, Italy*

Abstract

We propose a Deep Learning (DL)-based approach for production performance forecasting in fresh products packaging. On the one hand, this is a very demanding scenario where high throughput is mandatory; on the other, due to strict hygiene requirements, unexpected downtime caused by packaging machines can lead to huge product waste. Thus, our aim is predicting future values of key performance indexes such as Machine Mechanical Efficiency (MME) and Overall Equipment Effectiveness (OEE). We address this problem by leveraging DL-based approaches and historical production performance data related to measurements, warnings and alarms. Different architectures and prediction horizons are analyzed and compared to identify the most robust and effective solutions. We provide experimental results on a real industrial case, showing advantages with respect to current policies implemented by the industrial partner both in terms of forecasting accuracy and maintenance costs. The proposed architecture is shown to be effective on a real case study and it enables the development of predictive services in the area of Predictive Maintenance and Quality Monitoring for packaging equipment providers.

Keywords: Data Mining; Deep Learning; Equipment Provider, Food Industry, Industry 4.0, Overall Equipment Effectiveness, Packaging

* Corresponding author. Tel.: +39-049-8277760.
** This work has been supported by Regione Veneto project PreMANI (MANIFATTURA PREDITTIVA: progettazione, sviluppo e implementazione di soluzioni di Digital Manufacturing per la previsione della Qualità e la Manutenzione Intelligiente) and Regione Veneto project ADAPT4.0 (Sviluppo di un Sistema adattativo per il riconoscimento di guasti nell'Industria 4.0).
E-mail address: gianantonio.susto@dei.unipd.it

10.1016/j.promfg.2020.01.033

1. Introduction

Fresh product packaging is a highly demanding sector, where production performance forecasting is pivotal. On the one hand, this is a very demanding scenario where high throughput is mandatory; on the other, due to strict hygiene requirements, unexpected down time caused by packaging machines can lead to huge product waste. In fact, if machine breaks are not anticipated/predicted, the timing to fix the equipment before products have to be discarded is very limited when dealing with fresh products like dairy products and juices.

In this context, Predictive Maintenance (PdM) technologies [4, 21] can be extremely valuable. PdM is a particularly relevant technology for cost reduction and production optimization in the Industry 4.0 scenario [25]. Indeed, PdM solutions provide estimations of the health status of a machine [23], an information that can be exploited for taking optimized maintenance/service actions and production/logistics-related decisions [27].

Generally, PdM solutions are developed with the usage of Machine Learning approaches and by exploiting the availability of historical data [24]. For equipment providers, whose service-related challenges are complicated by the logistics associated with the geographical distribution of the installed machines, PdM is even more relevant. Indeed, faults affecting machines installed in places that are difficult to reach require long intervention time that may translate into huge product loss.

In this work, we employ Deep Learning-based forecasting techniques to predict production performance. In particular, our aim is predicting future values of two key performance indexes:

- **Machine Mechanical Efficiency (MME):** MME is the ratio between effective running time (ERT) and general running time (GRT), i.e. the time associated to production phase, deprived of down-time not associated to the machine. ERT and GRT are defined in the context of beverage packaging [6].
- **Overall Equipment Effectiveness (OEE):** Ideally, an OOE of 100% means that the production process is flawless, i.e. it is manufacturing only good parts, as fast as possible, with no downtime. Indeed, OOE is calculated as follows [14]:

$$OEE = Q \times P \times A.$$

where

  - **Q (Quality)** is the ratio of good parts over total parts. i.e. $Q = 1 - \frac{WA}{WA+GO}$, where WA and GO stand for waste and good packages, respectively;
  - **P (Performance)** is the ratio of actual speed over nominal speed, i.e. $P = 1 - \frac{NS-AS}{NS}$, where NS and AS stand for nominal and actual speed, respectively. In packaging industry, AS is typically defined in packages per hour. Notice that, in the computation of actual speed, we consider both good and waste packages.
  - **A (Availability)** is the ratio of total run time over planned production time [17]. i.e. $A = 1 - \frac{AAT+DTM+DTNM}{BWT}$ where AT is ancillary time, DTM is internal downtime and DTNM is external downtime.

2. Main Contribution

In this work, we address the problem of forecasting OEE and MME by leveraging Deep Learning-based approaches. Indeed, not only is Deep Learning (DL) capable of modeling complex, possibly non-linear relations among data sources, DL approaches can also automatically extract relevant features to solve the task at hand. In the proposed approach, we first design a model that focus only on historical production performance data. Then, we also include measurements about alarms occurred during production stage and combine them as a composite input to the neural network. This situation, where equipment data are monitored on-line and are able of sharing information about their current status, is often encountered in Industry 4.0 scenarios [16, 22]. In this work we show how it can be successfully used to improve prediction quality.

In particular, the proposed production performance forecasting solution is designed to be part of a Decision Support System (DSS). From the perspective of the equipment provider, the developed DSS enables smart monitoring

of deployed machines and it can help service managers in planning maintenance interventions; moreover, smart monitoring capabilities can be shared with customers to offer a sort of augmented service, where requirements coming from production orders and forecasted values for MME and OEE are collected to optimize production planning. In this work, experimental results on a real industrial case are obtained. Different architectures and prediction horizons are analyzed and compared to identify the most robust and effective solutions, showing advantages w.r.t. current policies implemented by the industrial partners both in terms of forecasting accuracy and maintenance costs. Finally, we remark that the present is one of the first work that exploit Machine Learning-based approaches to optimize packaging machines or service-related actions [1].

3. Deep Learning for Time Series

DL architectures [8] have been successfully applied in time series modeling and prediction tasks [7] with a broad variety of application fields, spreading from Natural Language Processing [26] to Anomaly Detection [15, 3]. The introduction of Recurrent Neural Networks (RNN) allowed to overcome the issues of traditional feed-forward neural networks in modeling sequences [10]. However, due to the well-known vanishing gradients problem of RNN [11, 19], recently Long Short Term Memory (LSTM) networks [12] have become more and more popular for sequence [5] and sequence-to-sequence [26] learning.

The main module of a LSTM can be seen as an ensemble of switch gates that enforce constant error flow through the internal states of special units called 'memory cells'. In this way, LSTM can learn nonlinear dynamics and model long-term dependencies better than RNN. However, recent studies have shown that also LSTM have some drawbacks [2]. In particular, their capability of learning long-term dependencies usually degrades for sequences longer than 100 elements. Moreover, LSTM training is particularly cumbersome from the point of view of computational memory.

Given the aforementioned problems in dealing with prediction performance forecasting, we compare LSTM with a recently developed architecture called Temporal Convolutional Network (TCN) [13]. The main advantages of a TCN over recurrent network, as a LSTM, can be summarized as follows:

- A TCN can "remember" long sequences, while, as already mentioned, the LSTMs have limited memory in time, in practice;
- A TCN do not suffer from the problem of vanishing gradient such as classic RNNs;
- A TCN needs less computational memory to be trained and the execution can be parallelized.

A brief introduction on TCNs is reported below, however we refer the interested readers to [2] for more details.

3.1. Details on Temporal Convolutional Networks

The experimental results in [2] suggest that TCN models using fully-convolutional network (FCN) architecture [20], dilated causal convolutions (as in Wavenet, [18]) and residual blocks [9], substantially outperform generic recur- rent architectures, as they are can be trained more easily and exhibit longer memory than recurrent architectures with the same capacity.

In TCN, convolutions must be causal to ensure that the prediction $P(x_{t+1}|x_1, \ldots, x_t)$ emitted by the model at time step t will not depend on any of the future time steps $x_{t+1}, x_{t+2}, \ldots, x_{t+n}$. A dilated convolution is a convolution where the filter is applied over an area larger than its length by skipping input values with a certain step. It is equivalent to a convolution with a larger filter derived from the original filter by dilating it with zeros, but is significantly more efficient. A dilated convolution effectively allows the network to operate on a coarser scale than standard convolution (see Fig. 1). Stacking dilated convolutions allows the network to have larger receptive field even with a few layers, while preserving the input resolution and computational efficiency.

The size of the convolution kernel may vary. The wider the kernel, the more input elements are taken into account at the first dilation and consequently at all subsequent dilations (see Fig. 2). TCN can "remember" long sequences, simply by increasing the number of dilations or the size of the kernel. We refer the reader to [9, 28] where it is demonstrated that residual blocks contribute to resolve the degradation problem in deep neural networks.

3.2. Regression loss Functions

In the equipment provider IoT scenario at hand, a major issue is the presence of outliers in acquired measurements, mainly due data inconsistency associated with communication issues and non-standard maintenance interventions. Given this premise, we focused on choosing an appropriate loss function for the TCN to robustly deal with outliers. Mean Square Error (MSE) is the most commonly used regression loss function. MSE is the sum of squared distances between our target variable and predicted values. Mean Absolute Error (MAE) is another loss function used for regression models. MAE is the sum of absolute differences between our target and predicted variables. So it measures the average magnitude of errors in a set of predictions, without considering their directions. When an error $e = y_{true} - y_{predicted}$ is greater than 1, the MSE, being the error squared (e^2) will obviously results in a value even bigger. This means that a model with MSE loss will give more weight to the outliers than a model with a MAE loss. This results in a model that is severely influenced by outliers, to the detriment overall prediction performance. However, the use of the MAE is far from being flawless, too, as it has constant gradient and therefore is affected equally by small and large errors. Thus, a better solution when dealing with outliers is to use $logcosh$, the logarithm of the hyperbolic cosine of the prediction error. It can be approximated as $e^2/2$ when the error is small and as $abs(e) - \log(2)$ when the error grows. This means that it resembles the MSE for small errors, but it doesn't weight too much the outliers (see Fig. 3), as it resembles MAE for bigger errors.

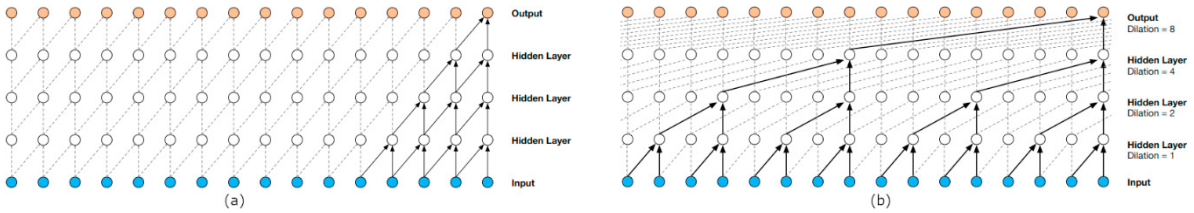

Fig. 1: (a) Stack of casual convolutions; (b) Stack of causal dilated convolutions. (Adapted from [18])

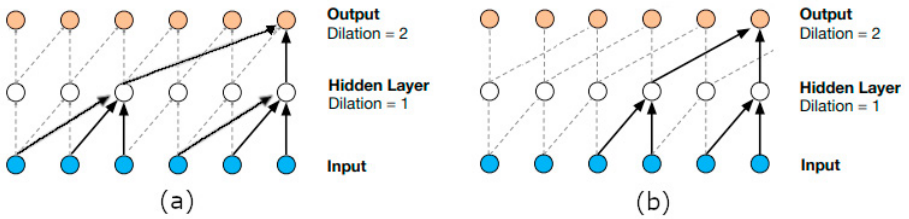

Fig. 2: Two dilated convolutions with different kernel size. (a) kernel size = 3, the output is derived from 6 initial input time steps; (b) kernel size= 2, the output is derived from 4 initial input time steps. (Adapted from [18])

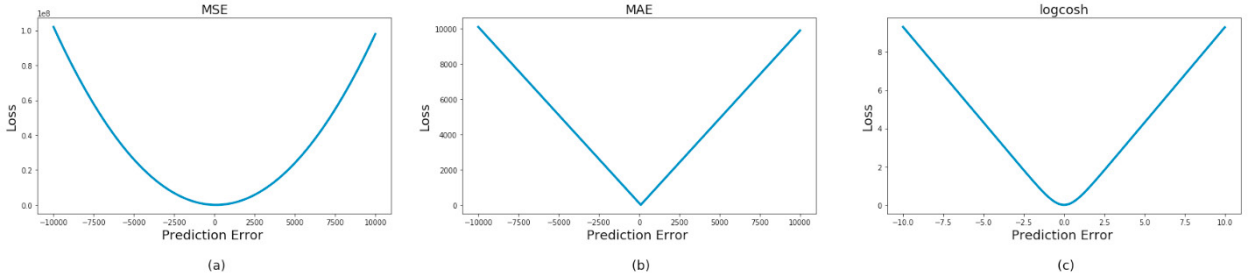

Fig. 3: Three different loss functions considered: (a) MSE; (b) MAE; (c) logcosh. It can be seen that, while MSE is a universally adopted metric, logcosh has a clear advantage in the presence of outliers since the effect of larger prediction errors is mitigated.

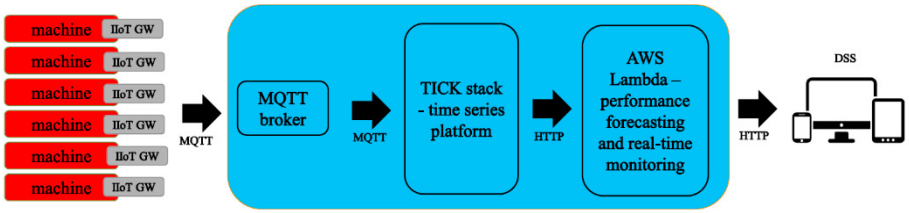

Fig. 4: IoT architecture. Acronyms and technologies: Industrial Internet-of-Things Gateway (IIoT GW); Message Queue Telemetry Transport (MQTT); TICK Stack is a collection of associated technologies which combine to deliver a platform for storing, capturing, monitoring and visualizing data that is in time series; AWS Lambda is an event-driven, serverless computing platform provided by Amazon.

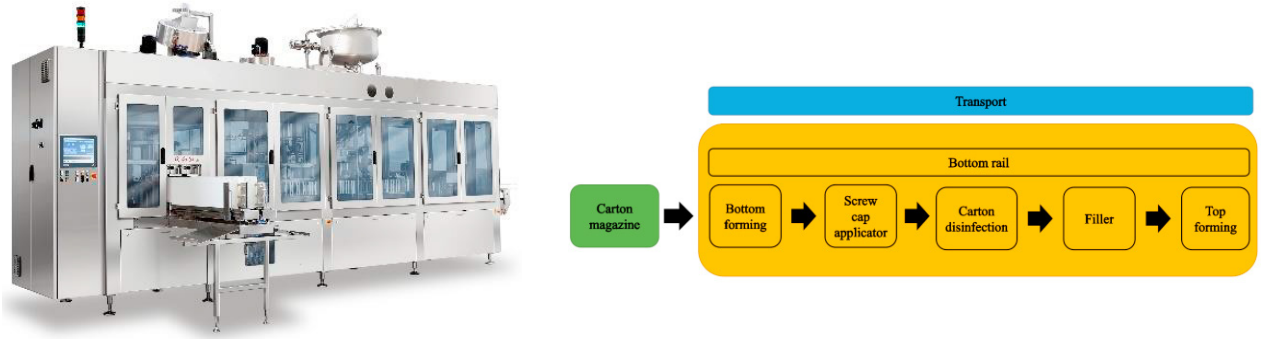

Fig. 5: Automatic filling machine analysed in this work: picture (left panel) and block scheme describing the main modules (right panel).

4. Use Case

We tested the proposed approach in a real industrial scenario, where pieces of packaging equipment are monitored in real-time by means of an IoT framework that acquires information about current production stage, such as process settings, warnings and alarms generated by the equipment. Recently, the industrial partner has developed a cloud- based infrastructure, described in Fig. 4, and in this analysis we are considering data related to 6 machines with approximately 18 months of data for each piece of equipment.

The object of our study is an automatic filling machine model designed for inline forming, filling and sealing of gable top cartons, depicted in Fig. 5 (left panel). The machine is composed by the following main subsystems, depicted in Fig. 5 (right panel): cartons magazine, bottom rail, transport, bottom forming, screw cap applicator, carton disinfection, filler and top forming.

We can measure the time intervals when a machine has been in production, cleaning or downtime status. Moreover, we monitor 640 different types of alarms that can be divided into two classes:

- *Internal alarms* caused by malfunction in the equipment;
- *External alarms* generated due to errors in how the equipment is used by human operators.

Even if the machines are monitored in real-time, our purpose was to predict the single measure (MME or OEE) day by day; so, we aggregated the data daily. Therefore, the dataset used to train the model has the following information with a daily granularity: MME, OEE, Number of Internal Alarms, Number of External Alarms, Production time, Cleaning time, Down time. Note that the MME and OEE, being measurements, are averaged in the aggregation, while the number of alarms and the time intervals on the machine status have been summed up.

Since MME and OEE had different trends over time, we decided to use two different models. Therefore, the model for the prediction of the MME uses all the columns of the data set except the one related to the OEE and viceversa. For both models we have decided to use an architecture based on a TCN, where the residual-block is composed of 2 convolutional layers. We found through cross-validation that for the specific case (i.e. predicting the next day) choosing an input window too wide would introduce noise and thus worsen the results. Our choice fell then on a window of two weeks. Cross-validation allowed us to identify as 2 the ideal size of the convolutional

kernel. With this formula (which takes into account the fact that the residue block is composed of 2 convolutional layers) it's possible to calculate the Receptive Field of our proposed network:

$$ReceptiveField = 1 + 2(k_s - 1)(2^n - 1),$$

where k_s is the kernel size and n is the number of dilations. So, with the size of the kernel equal to 2, the minimum number of causal dilations that allows us to take into account all previous 14 days is 4.

Since we used a real data set, it brought with it all the problems that follow. In fact, the IoT framework that captures information from machines is not lossless. In addition, the MME and OEE metrics by their very nature have a tendency to have outliers (as depicted in Fig. 6). For example, the average MME of a day for a car has a constant value of around 80, but on days when there are problems it can go down to 0. This behavior can be observed in Fig. 4. For this reason, as discussed in previous section, we decided to use $logcosh$ as a loss function more robust than MSE to outliers and more suitable than MAE to gradient-based optimization.

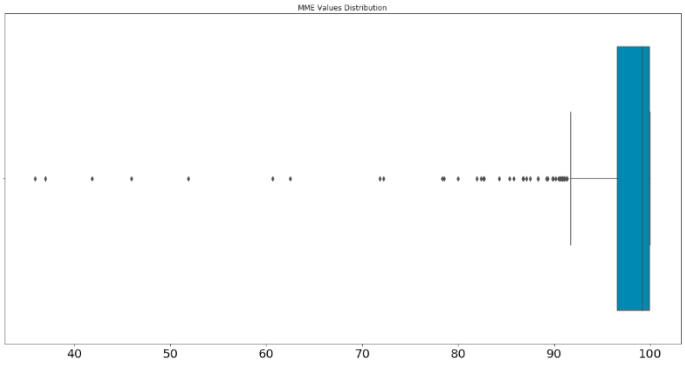

Fig. 6: Distribution of the average values per day of MME for a specific machine. It's clearly visible that there are many outliers. These outliers may be the result of a loss of data in the acquisition framework, or they may correspond to days when the machine had major problems.

5. Results

We compared TCN with a classic LSTM-based architecture.

The same tests were performed on both architectures by considering increasingly complex models. At first, we considered a model based only on historical production performance data to predict MME and OEE ('Experiment #1'). Then we added error count to the input ('Experiment #2'), and, finally, also information about machine production status are included ('Experiment #3'). The results are summarized in the Tab. 1.

The architecture of the models remains basically invariant, only the number of channels in the input signal varies according to available information sent to the network. Hyper-parameters are optimized for each configuration separately by means of grid-search.

For evaluating the results, we used the root mean squared error (RMSE) and performed cross validation over time. In particular, about half of the time frame is used as a training window and the following 2 weeks of data are considered for the test. Then, the two windows are shifted by two weeks in a rolling fashion. In this way, the size of the training dataset remains the same. Maintaining all the data of the historical time series for the training dataset (allowing the training data to increase over time) would lead to an improvement in the model performance over time. However, we are interested in evaluating the performance of the proposed approach in a less favorable scenario where each of the trained model has access to the same amount of information.

It is interesting to note from Tab. 1 that, except in the case with the dataset formed by metrics and error numbers, the TCN obtains better results than LSTM. Although the improvement is not drastic, we see that the addition of features leads to an improvement in results. In fact, in no case a data set with fewer features get better scores than its enriched counterparts.

Table 1. Experimental results of a specific machine with data over a period of more than one year, precisely 400 days. The machine chosen to show the results was the one with the least loss of information.

Dataset (Experiment #)	Model	Metric	RMSE
Only Metric (Experiment #1)	TCN	MME	**18.04**
		OEE	**16.05**
	LSTM	MME	19.60
		OEE	17.26
Metrics + errors counts (Experiment #2)	TCN	MME	**15.73**
		OEE	19.78
	LSTM	MME	17.74
		OEE	**17.11**
Metrics + errors counts + status intervals (Experiment #3)	TCN	MME	**12.66**
		OEE	**14.89**
	LSTM	MME	14.79
		OEE	17.41

6. Conclusion

In this work we have presented a Deep Learning-based approach to production performance prediction for industrial machines; in particular, we have validated our approach on the prediction of Machine Mechanical Efficiency and Overall Equipment Efficiency for fresh product packaging equipment. We remark that the availability of efficiency prediction can be exploited both by the equipment provider, to optimize maintenance strategies, and by end users (production line owners) that can exploit prediction to estimate future throughput; this is particularly relevant when dealing with fresh products, where unexpected downtime can easily lead to product waste. The experiments reported here are promising w.r.t. important goals for equipment providers like cost minimization and increased sustainability. Equipment providers are nowadays implementing IoT-based solutions to remotely monitor their equipment installed in the production lines of their customers and they are developing automatic tools for descriptive data analysis; in future scenarios, the proposed Deep Learning solution can enable the transition to predictive analytics. Another important impact of the work detailed in this manuscript is the acquired know-how achieved by analyzing the available data: the gained insight on data quality have been exploited to enrich the acquisition system and its scope. In particular, the acquisition framework has been redesigned to include the most relevant sensor data to allow a more fine-grained understanding of equipment status.

Some future research directions are foreseen:

- Extension of the preliminary results reported here on a large scale dataset with dozens of machine installed worldwide;
- Given the availability of sensor data, the development of Predictive Maintenance solutions for equipment sub-systems whose predictions can be exploited by the production forecasting model for improved accuracy.

We underly that the architecture proposed in this work was shown to be effective on a real case study, enabling the development of predictive services in the area of PdM and Quality Monitoring for packaging equipment providers.

References

[1] R. Accorsi, R. Manzini, P. Pascarella, M. Patella, S. Sassi. Data mining and machine learning for condition- based maintenance. *Procedia*

Manufacturing, 11:1153–1161, 2017.

[2] S. Bai, J. Kolter, V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv:1803.01271*, 2018.

[3] S. Chauhan, L. Vig. Anomaly detection in ecg time signals via deep long short-term memory networks. *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–7, 2015.

[4] A. Chen, J. Blue. Recipe-independent indicator for tool health diagnosis and predictive maintenance. *IEEE Transactions on Semiconductor Manufacturing*, 22(4):522–535, 2009.

[5] A. Dai, Q. Le. Semi-supervised sequence learning. In *Advances in neural inform. processing systems*, 3079–3087, 2015.

[6] Deutsches Institut fur Normung E.V. (DIN). Beverage packaging technology; terminology associated with filling plants and their constituent machines. 1984.

[7] F. Gers, D. Eck, J. Schmidhuber. Applying lstm to time series predictable through time-window approaches. In *ICANN*, 2001.

[8] I. Goodfellow, Y. Bengio, A.Courville. *Deeplearning*, volume1. MIT press Cambridge, 2016.

[9] K. He, X. Zhang, S. Ren, J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] M. Hermans, B. Schrauwen. Training and analysing deep recurrent neural networks. In *NIPS*, 2013.

[11] S. Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998.

[12] S. Hochreiter J. Schmidhuber. Longshort-termmemory. *Neuralcomputation*,9(8):1735–1780,1997.

[13] C. Lea, R. Vidal, A. Reiter, G. Hager. Temporal convolutional networks: A unified approach to action segmentation. In *European Conference on Computer Vision*, pages 47–54. Springer, 2016.

[14] D. Liao, W. Tsai, H. Chen, Y. Ting, C. Chen,H. Chen, S. Chang. Recurrent reinforcement learning for predictive overall equipment effectiveness. In *2018 e-Manufacturing & Design Collaboration Symposium (eMDC)*.

[15] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, G. Shroff. Lstm-based encoder-decoder for multi-sensor anomaly detection. *CoRR*, abs/1607.00148, 2016.

[16] J. Moyne, J. Samantaray, M. Armacost. Big data capabilities applied to semiconductor manufacturing advanced process control. *IEEE transactions on semiconductor manufacturing*, 29(4):p283–291, 2016.

[17] P. Muchiri, L. Pintelon. Performance measurement using overall equipment effectiveness (oee): literature review and practical application discussion. *International journal of production research*, 46(13):3517–3535, 2008.

[18] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[19] R. Pascanu, T. Mikolov, Y. Bengio. On the difficulty of training recurrent neural networks. In *ICML*, 2013.

[20] E. Shelhamer, J. Long, T. Darrell. Fully convolutional networks for semantic segmentation. *arXiv preprint arXiv:1605.06211*, 2016.

[21] G.A. Susto, A. Beghi. Dealing with time series data in predictive maintenance problems. In *Emerging Technologies and Factory Automation (ETFA), 2016 IEEE 21st International Conference on*, pages 1–4. IEEE, 2016.

[22] G.A. Susto, A.B. Johnston, P. O'Hara, S. McLoone. Virtual metrology enabled early stage prediction for enhanced control of multi-stage fabrication processes. In *Automation Science and Engineering (CASE), 2013 IEEE International Conference on*, pages 201–206. IEEE, 2013.

[23] G.A. Susto, A. Schirru, S. Pampuri, A. Beghi, G. De Nicolao. A hidden-gamma model-based filtering and prediction approach for monotonic health factors in manufacturing. *Control Engineering Practice*, 74:84–94, 2018.

[24] G.A. Susto, A. Schirru, S. Pampuri, D. Pagano, S. McLoone, and Alessandro Beghi. A predictive maintenance system for integral type faults based on support vector machines: An application to ion implantation. In *Automation Science and Engineering (CASE), 2013 IEEE* p. 195–200.

[25] G.A. Susto, J. Wan, S. Pampuri, M. Zanon, A.B. Johnston, P. O'Hara, S. McLoone. An adaptive machine learning decision system for flexible predictive maintenance. In *Automation Science and Engineering (CASE), 2014 IEEE International Conference on*, pages 806–811.

[26] I. Sutskever, O. Vinyals, Q. Le. Sequence to sequence learning with neural networks. In *NIPS*, pg 3104–3112, 2014.

[27] J. Wang, L. Zhang, L. Duan, R. Gao. A new paradigm of cloud-based predictive maintenance for intelligent manufacturing. *Journal of Intelligent Manufacturing*, 28(5):1125–1137, 2017.

[28] S. Zagoruyko, N. Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*,2016.