Forecasting cycle time in semiconductor manufacturing systems: a literature review

Masi A.*, Pero M.*, Cannas V.G., Ciccullo F.***

** Dipartimento di Ingegneria Gestionale, Politecnico di Milano, Piazza Leonardo da Vinci, 32 20133 - Milano - Italy (antonio.masi@mail.polimi.it, margherita.pero@polimi.it, federica.ciccullo@polimi.it)*

*** Scuola di Ingegneria, Università Carlo Cattaneo – LIUC, Corso Matteotti 22, 21053 - Castellanza – Italy (vcannas@liuc.it)*

Abstract: An efficient and effective forecasting of production cycle times (CT) is a critical success factor in semiconductor manufacturing systems (SMS): inaccurate CT forecasts can have a negative impact on production scheduling, causing late deliveries, as well as on the amount of inventories and work-in-progress, which rapidly lose value over time because of the high risk of obsolescence. Therefore, since the 80s, several quantitative techniques have been developed to face this problem. Furthermore, Artificial Intelligence (AI) techniques are gaining importance, despite their potential is still not fully exploited even in the most advanced manufacturing systems. However, a synthetic overview of the techniques to forecast CT in SMS is still missing in the literature. As a result, it is difficult for decision makers to orient themselves and choose, among the many existing ones, the best model for their specific situation, comparing the different performance in terms of accuracy, data required, speed and easiness to use. This paper aims at presenting an overview of the quantitative techniques developed to forecast production CT in SMS. Firstly, a description of the methodology with which the literature review has been carried out is provided. Secondly, a taxonomy of forecasting techniques is proposed. Subsequently, a synthetic description of analytical, simulation, time-series and causal methods is presented. Within statistical techniques, a special focus is deserved to AI ones, since their popularity has dramatically increased in the last years. In particular, the most recent applications of artificial neural networks (ANN) in SMS – namely, hybrid methods and Long-Short-Term-Memory recursive neural networks – are described. Finally, a table with a qualitative comparison between the different methods is proposed.

Keywords: forecasting, cycle time, semiconductors, artificial intelligence, neural networks

1. Introduction

Semiconductor manufacturers are key players within electronic products supply chain. Global semiconductor industry sales reached US$468.8 billion in 2018 with an increase of 13.7% compared to the 2017 (according to the Semiconductor Industry Association). Semiconductor manufacturers produce integrated circuits on silicon wafers. The wafer production is the most complicated, expensive and time consuming phase of Integrated Circuits manufacturing (Connors et al., 1996). Wafers are produced in plants called "fabs". Fabs are large manufacturing facilities, with tenths to hundreds of machines moving hundreds to thousands of wafers (Mönch et al., 2011).

Planning and managing fabs is complex for various reasons: (i) each lot of wafers has to perform a specific set of different operations in a different order, (ii) each operation can be performed by different machines, (iii) a machine can perform different operations, and (iv) there are lots that must perform more than once a specific operation in the same machine (Dabbas and Fowler 2003).

Among the planning activities, cycle time (CT) forecasting, i.e. forecasting the time from when a lot of wafers is released into the fab until it is ready as a finished product (Kumar and Kumar, 2001), is a challenging problem, but semiconductor manufacturers need to have tools to solve it. In fact, having a good estimation of CT improves planning performance, by reducing overstock or late deliveries. Overstock results in high inventory costs due to the high perishability (wafers are sensitive and should not be exposed for long times to aerial contaminants between production steps) and risk of obsolescence of the wafers. Secondly, since wafers are mostly produced based on orders, it is fundamental to be able to provide customers with a reliable forecast of when their wafers will be ready for shipment.

Different methods have been developed over the years to tackle the problem of CT forecasting, and new methods, connected to Artificial Intelligence (AI) development, have been proposed. However, to the best of the authors knowledge, there is a lack of literature reviews covering this topic. Therefore, the research objective (RO) of this paper is to systematically analyse the literature on the available methods for CT forecasting in semiconductor manufacturing systems (SMS), including new trends (e.g. AI-based methods) to provide a comprehensive overview for both researchers and practitioners and to compare the available approaches according to relevant performance dimensions.

The remainder of the paper is organized as follows: in section 2 the methodology used for the literature review is described along with the descriptive analysis of the papers included; then, section 3 presents the classification of the different methods, their description and their comparison; section 4 presents the main research gaps and future trends;

finally, section 5 is devoted to conclusions and discussion of this study limitations.

2. **Literature review**

In this study, a rigorous and systematic literature review about the CT forecasting techniques in SMS was conducted. According to a methodology similar to the one adopted by Cicculo et al. (2017) and Gosling and Naim (2009), the work was organised along three steps:

- material collection (section 2.1), which describes how the articles were selected;
- descriptive analysis (section 2.2), which provides a temporal and journal distribution of the papers found, as well as a description of the type of data considered (real or simulated);
- results, which presents the key findings of the thematic analysis of the research (section 3).

2.1 Material collection

The material collection was articulated in two steps (see Figure 1): firstly, an initial set of papers was identified through a structured query, and it was subsequently filtered by examining the titles and the abstracts to check the pertinence to the research area; secondly, also the full bodies of the articles were analysed, in order to discard the works which were not close to the specific research question. The main data source used for the research was Scopus, the largest abstract and citation database of peer-reviewed literature.

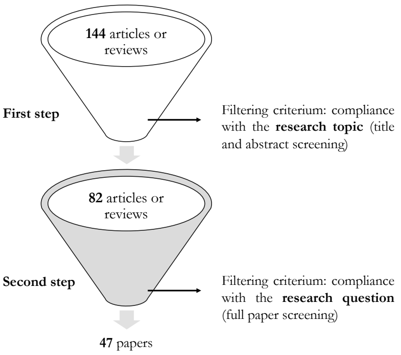
144 articles or reviews
First step
Filtering criterium: compliance with the research topic (title and abstract screening)
82 articles or reviews
Second step
Filtering criterium: compliance with the research question (full paper screening)
47 papers

Figure 1: material collection

A critical part of the first step was the definition of a proper query. "Forecast*", "cycle time" and "semiconductor" were the keywords which were closest to the research topic. After carrying out other exploratory researches and trying different queries, the main variants of these first three keywords were identified. Eventually, a query composed by three keyword sets was constructed:

TITLE-ABS-KEY (("cycle time" OR "output time" OR "flow time" OR "due date" OR "completion time") AND (predict* OR forecast*) AND (wafer OR semiconductor)) AND DOCTYPE (ar OR re)

As a starting point, the research was restricted only to articles or reviews; in the future, it will be possible to extend it also to conference papers, in order to have an even broader view on the subject and to include the most innovative models.

Out of 144 potential papers, 82 were selected after the first step. The discarded articles were mainly works more related to technological aspects of SMS than managerial ones; for example, articles related to motions of the quantum particles in semiconductor circuits.
After the second filter, 47 articles were eventually selected. The discarded articles were belonging to the Operations Management area, but not to CT forecasting one. In this group, a noteworthy category was represented by the papers concerning dispatching and other scheduling issues. The Scopus database was searched in different time slots from November 2018 to February 2019.

2.2 Descriptive analysis

The selected articles cover a time from 1999 to 2019, as it is shown in Figure 2. Their distribution over time is mainly concentred between 2006 and 2009; however, the interest towards thus subject is still remarkable, as the peak in 2018 shows.

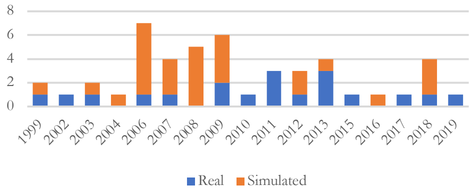

Figure 2: temporal distribution of papers in the final sample, divided according to the type of data analysed

The selected articles come from a set of 26 different sources. In particular, about the half of the papers belongs to the 20% of the sources, which are five top rank (Q1) journals in their field: 7 papers to *International Journal of Advanced Manufacturing Technology* (SJR 2017 0.99), 5 papers to the *International Journal of Production Research* (SJR 2017 1.47), 4 papers to the *Journal of Intelligent Manufacturing* (SJR 2017 1.18), 4 papers to the *Applied Soft Computing Journal* (SJR 2017 1.2), and 3 papers to *Computers and Industrial Engineering* (SJR 2017 1.46). Attention was given also to the number of citations, the average number of citations that the articles selected received is 30 citations, and 21 articles out of 47 are included in a range between 30 and 119 citations.
All the articles have an experimental part, in which the proposed models are tested. Two approaches are possible: using real data, coming from an actually existing plant, or simulating them with a software. Articles based on simulated data represent the 60% of the works included in the corpus. However, as Figure 2 shows, the interest for works with real data slightly increased after 2010.
38 articles reported the number of jobs considered in their datasets, while the remaining ones were only stating the time horizon analysed (e.g. "three months of production"). As far as the former group is concerned, it is possible to prove that simulated datasets are usually larger than real ones, as Figure 3 and Figure 4 show. However, it is worth mentioning that two simplifying assumptions were made in this analysis: firstly, in the few cases in which different datasets were considered in an article, only the largest dataset was taken into account for its allocation in the

histogram. Secondly, two "outliers" have not been represented, namely the papers by Chen et al. (2009a), which simulated a dataset of 8,700 jobs, and by Tirkel (2013), which started with a set of 8,000 real observations. Although the latter dataset was reduced to about 6,865 records after the cleaning phase, it still represents the largest study conducted on real data among the examined papers.

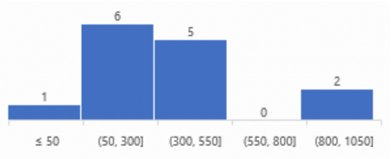

Figure 3: histogram of the size of the real datasets

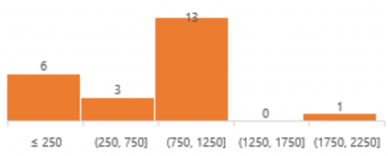

Figure 4: histogram of the size of the simulated datasets

3. Results

The analysis of the bodies of the selected papers led to two main results: (i) an updated taxonomy of Cycle Time Forecasting (CTF) (sections 3.1 and 3.2), and (ii) a qualitative comparison (section 3.3).

In the last twenty years, different classifications of CTF methods have been proposed, ever since Huang et al. (1999) identified as main forecasting techniques "simulation, queueing, spreadsheet, regression and neural networks". This study was based on two main models: Wang and Zhang (2016) and Chen (2007a) - as a further developed of the one of Chen (2003) - integrated, modified and expanded with contributions coming from other papers examined during the review. To integrate the two perspectives, we distinguished between the "primary techniques" and the "derived techniques", which combine the former ones in various ways.

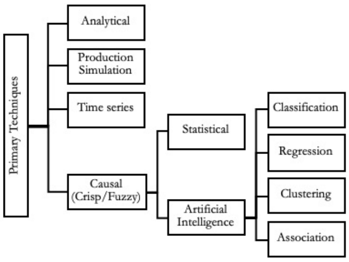

Figure 5: taxonomy of the primary CTF techniques

3.1 Primary techniques

As far as primary techniques are concerned, the following classification of these techniques is proposed: (1) analytical methods; (2) time series methods; (3) causal methods; (4) production simulations. A breakdown of these classes is discussed in the following paragraphs and summarized in Figure 5.

3.1.1 Analytical methods

Analytical methods try to model the functioning of a wafer fab by picturing it with a sort of flow chart and using mathematical relationships to model variables like arrivals, lead times and saturation of the machines. As stated by Wang and Zhang (2016), these methods include: Queuing Theory (QT), Markov chain and Petri net.

Analytical methods have been extensively applied in the case of uncomplicated situations; for instance, if only two priority classes are present. In terms of accuracy, Kuo et al. (2011) obtained a lower Mean Absolute Percentage Error (MAPE) with artificial neural networks (ANN) (7.6% on average) than with a queuing model (28.9% on average) on a dataset of 502 samples.

3.1.2 Production simulations (PS)

PS create virtual models of wafer fabs in order to study their behaviour and predict interesting variables, and CT is one of them. One of the most commonly employed PS techniques is Discrete Event Simulation (DES), which is mentioned by Wang and Zhang (2016).

PS are widely used in research and development to simulate data to test other models, as it was stated in section 2.2. When it comes to CT forecasting, even though they can achieve among the highest levels of accuracy, they present several drawbacks. Chen (2007b) mentioned some of the shortages of PS: huge amount of input data required; need to continuously update such databases; long simulation time, "with thousands of replications necessary to sufficiently consider all the uncertain events". More recently, Kuo et al. (2011) observed that "the construction and implementation of a simulation requires significant effort and lengthy computer time. Furthermore, since the modelling of a simulation is based on a specific system, it can hardly serve as a generic solution".

3.1.3 Time series methods

In time series methods, as it is explained by Tirkel (2013), the forecast of the CT of a job is based only on the CT of the previous ones, which represent the time series to be analysed. This category of methods contains techniques such as:

- Autoregressive Integrated Moving Average (ARIMA) models, with their variants Kohonen-ARIMA (KARIMA) and Seasonal ARIMA (SARIMA), (Tin et al. 2019);
- Moving Average (MA) and Weighted Moving Average (WMA);
- Exponential Smoothing (ES) – which is employed, for example, in MFLC (see section 3.2.1).

In these methods, as stated by Chen and Romanowski (2013), "the cycle time of a job is usually predicted when the job is released into the factory, but several months are required for the job to complete all operations". The need to make a several steps-ahead prediction is one of the reasons for the low accuracy and low diffusion of these methods.

3.1.4 Causal methods

Causal methods follow an input-output relationship approach – to use the terminology of Chen and

Romanowski (2013) - "to determine certain factors (e.g., average waiting time, queuing length, utilization, future release plans, bottlenecks) that influence the job cycle time, and then to apply different approaches, e.g. multiple linear regression (MLR) and ANN, to model the relationship between the job cycle time and these factors, in order to forecast the cycle time of a new job". Within this family it is useful to separate "traditional" statistical methods, like MLR, from AI methods, which bloomed after the forth industrial revolution. In fact, AI methods need large datasets to be used as "training data" and require heavier computations, but they can improve by orders of magnitude the accuracy respect to statistical ones. For example, Chen and Romanowski (2013) obtained a Mean Absolute Percentage Error (MAPE) of 0.9% with a hybrid AI method, against the 6.1% of MLR.

A second distinction should be done according to the "logic" adopted by the model, which can be "crisp" or "fuzzy":

- the crisp approach follows a binary logic, according to which a statement can have only two possible states (e.g. 1/0, "true"/ "false");
- the fuzzy approach, which is the one on the basis of soft computing, follows a fuzzy logic, according to which the same statement can assume a range of possible statements (e.g. "low", "medium", "high"). As a consequence, when operating with sets of items, the same item can belong to different sets with different degrees of confidence.

A model belonging to a certain family can be used either with a crisp or with a fuzzy approach, for example one can use MLR or its fuzzy variant, Fuzzy Multiple Linear Regression (FMLR).

Within statistical methods, it is possible to include:

- regression based methods (Wang and Zhang 2016), like LR, MLR, and its already mentioned fuzzy variant – FMLR – employed in Chen (2009);
- probability distribution-based methods, like Weibull-distributed fitting (Wang et al. 2018);
- relation analysis (Sha and Hsu 2004).

As observed by Wang and Zhang, "since historical conditions are used to forecast the future, these statistical methods have poor accuracy in dynamic manufacturing systems". Moreover, in different studies by Chen MFLR performed poorly – for example, in Chen (2011), "the coefficient of determination (R2) was only 0.39, which revealed that the cycle time of jobs were highly uncertain and very difficult to estimate" with a linear relationship model. However, the speed of the computations and the high interpretability of the results make these models still widespread, particularly when applied to estimate the weights of MFLC (see section 3.2.1).

As far as AI methods are concerned, it is possible to make a further distinction (Kuo et al. (2011) between four tasks: classification; regression; association; clustering.

Regression tasks are very common in the field of CTF in SMS. However, also classification tasks are widespread. For example, understanding in which range of CT (e.g. "1 month", "2 months", "3 months") the next observation will fall is a typical classification task.

The most common algorithms in SMS for classification and regression (depending on whether the target variable – CT – was modelled) are Decision Trees (DT) and ANN.

DT - or Regression Trees (RT) – employed for instance in Backus et al. (2006), Hsu et al. (2006) and in Tirkel (2013) who, as mentioned in section 2.2, used them on a dataset of thousands of real records. Hence, speed and scalability are some of the strengths of these approaches. Moreover, if the resulting tree is not too complex, it is possible to visualize it, which makes the model more understandable and interpretable.

ANN, can be furtherly divided, according to Kuo et al. (2011), into:

- Back Propagation Networks (BPN) and their fuzzy variant Fuzzy Back Propagation Networks (FBPN);
- Radial-Basis Function Neural Networks (RBFN), employed in Wang et al. (2018);
- Recurrent Neural Networks (RNN), which include, among the others, Long Short-Term Memory (LSTM) networks (Tin et al. 2019).

BPN are the most widespread type of ANN. Out of 15 articles testing an ANN method (excluding Hybrid methods), 8 employ BPN.

As far as association rules are concerned, in the field of SMS Evolving Fuzzy Rules (EFR) are employed by Chang and Liao (2006) to produce rules of the type "if… then" by looking at recurrent patterns within the dataset. In this sense, CBR – employed by Chiu et al. (2003) and Chang et al. (2009) – can be also considered as a kind of associative algorithm.

Finally, the most popular clustering algorithms in CTF are K-Means (KM) (Chen 2007a) and especially its fuzzy variant Fuzzy C-Means (FCM), employed in 10 of the articles analysed.

3.2 Derived techniques

3.2.1 Multiple-Factor Linear Combination (MFLC)

Chen (2007a) explained that in MFLC "the cycle time of a lot is estimated with a weighted sum of parameters" including job properties (e.g. total processing time, number of re-entrances, number of operations), cycle time and waiting time series (e.g. "the numbers of operations of some (usually three) most recently completed lots"), and workload information (namely, Work-in-Progress levels). MFLC is the easiest, quickest and most prevalent in practical applications. In fact, several companies use it to estimate the internal due dates of the jobs. Some examples of these rules are:

- Total Work Content (TWK), where due dates are based on the total processing time;
- Number of Operations (NOP), where due dates are estimated according to the number of necessary operations;
- Jobs-in-Queue, where due dates are estimated on the basis of the current queue lengths in the system (JIQ) or in the bottleneck station (JIBQ).

MFLC can use different primary techniques as building blocks. In fact, the coefficients of the linear combinations need to be defined in other ways.

A first example is the exponential smoothing due date assignment rule, which employs the ES time series method:

$$CT_n = TP_n + D_n$$

where CT_n is the CT of lot *n*, TP_n is its total processing time, and D_n is a coefficient continuously updated as follows:

$$New\ D_n = D_n + \alpha(D_1 - D_n)$$

And α is a positive constant always lower than 1.

A second example is the TWK calculation as it is presented by Sha and Hsu (2004):

$$d_i = r_i + k\ \times p_i$$

where d_i denotes the assigned due date for the i-th order and *k* is the parameter that reflects the expected queue time which that order will experience in the system. The value of *k* is usually estimated via regression, which is a statistical method, but in the aforementioned paper it is estimated with a BPN, which is an AI method.

3.2.2 Hybrid methods

Hybrid methods combine different AI techniques, namely clustering with regression (or, sometimes, classification) ones. By doing so, they have shown to enhance the accuracy of regression algorithms. Intuitively, this happens because the regression algorithm is trained on a cleaner, more homogenous dataset; hence, during the learning process, the internal parameters of the algorithm are better estimated.

Chen (2007a) proposed an FCM-BPN where examples classified in different categories by the FCM are then learned with different BPNs, which are eventually combined in a further BPN ensemble that will be used to forecast the CT of the future jobs. The experimental results on simulated data showed a much lower Root Mean Squared Error (RMSE) with the FCM-BPN (42.98 h for the dataset B "hot" considered in this study) than with MFLC (84.17 h) and even BPN (81.53 h), thus showing a superior effectiveness. In terms of efficiency, it is worth remarking that, according to a table provided by the author, it took several hours to construct and train the FCM-BPN model - more than the time necessary to build a BPN one and much more than using MFLC. As far as computational speed is concerned, as Chen states, "the cycle time of a job in a semiconductor fabrication factory might be up to 3 months, and therefore forecasting taking only a few minutes is acceptable in practical applications". Chen and Lin (2009) developed an FCM-FBPN and tested it on real data. Similarly to the previously mentioned technique, the FCM is used for pre-classification and the FBPN for prediction. In this study a team of production experts was interviewed to assess the relative importance of the eight variables considered in the model and then these opinions were translated into fuzzy numbers. Again, the model outperformed MLFC, BPN and other old methodologies, achieving even half of the RMSE of MFLC in certain cases. Chen et al. (2009b) also proposed a post-classification technique. First, the CT of the jobs is predicted with a FBPN. As far as the FBPN is concerned, in this study also the different importance of the various experts interviewed is taken into account. Then, the average error on the training data is computed. Later, the jobs are divided in two families – high or low expected accuracy – and each family is assigned to a code. Finally, a BPN is used to post-classify the same jobs according to the two categories: in other words, the BPN identifies the relationships between the features of the job and the expected accuracy. As a result, when the model will be applied on the new jobs, they will first pass through the BPN to determine to which of the two families they belong, and, later, they will be processed according to an appropriate FBPN, "tailored" for that family. Once again, the experimental results showed the superior performance of the hybrid methods respect to primary ones. However, it is worth noticing that the proposed methodology achieved results just slightly better (always less than 24 h of RMSE) than a simpler KM-FBPN.

3.3 Performance comparison

In order to compare the different methods proposed, several metrics can be considered. Similarly to the dimensions considered in Chen (2008), accuracy, speed, data required and easiness to use will be discussed in this paper.

Despite the high variety of methods that fall within each of the families discussed in this work, some general trends can still be identified, and, on the basis of them, it is possible to assign qualitative marks. These marks are not absolute; instead, they aim to establish a ranking between the different techniques.

The accuracy dimension is the easiest to assess, since there are quantitative indicators like the MAPE and the RMSE, which have been mentioned (whenever possible) for each method in its corresponding section. Generally, the most accurate techniques are hybrid methods, followed by PS, AI, analytical models, MFLC, statistical, and time-series.

In terms of speed, it would be necessary to assess computation time and human effort. According to the information already cited in the aforementioned section, simulations are the slowest models to build and execute, followed by hybrid models, AI, analytical, MFLC, and finally statistical and time-series analysis.

PS require large and constantly updated databases, while hybrid models, when they adopt a fuzzy logic, need teams of experts to correctly assess the fuzzy intervals: therefore, these two families can be considered the most demanding in terms of data required. AI and statistical methods are both causal techniques and therefore apply the same type of data, such as information related to WIP levels, length of the queues, fab utilization rates and so on. Analytical methods consider parameters (e.g. machine utilization rate) which are similar to those of causal ones, but not at a single-job specific level. MFLC, which combines analytical elements with statistical ones, should have a similar ranking. Finally, time-series methods, which consider very high-level data (the time series past CT themselves), are the cheapest to build.

Easiness-to-use is the most subjective dimension to assess. Chen (2011) assumed hybrid approaches to be the most difficult ones, followed by AI models (especially when applied with a fuzzy logic) and, finally, by MFLC. As it was mentioned in section 3.1.1, analytical methods are usually used when dealing with uncomplicated fabs; hence, it seems reasonable to assume for them at least a medium level of complexity.

Table 1 summarizes the comparison of the CTF methods. Based on the abovementioned considerations, a 5-scale rating has been used to compare the metrics: "very low", "low", "medium", "high", "very high". Very low accuracy means MAPE>6%; whereas very high means MAPE<1%. Very low speed means model construction/learning time < 5 mins; whereas very high means model construction/learning time > 8 hours. Very low data required means that the model needs high-level data; whereas very high means that the model needs constantly updated databases. Very low easiness to use means that the method is not intuitive, and users face issues while using it; whereas very high means that the method is intuitive, and users do not face issues while using it.

Table 1: comparison of the CTF methods

Metrics	Analytical	PS	TS
Accuracy	Medium	Very High	Very Low
Speed	High	Very Low	Very High
Data required	Low	Very High	Very Low
Easiness to use	Medium	Medium	High

Metrics	Statistical	AI	MFLC	Hybrid
Accuracy	Low	High	Low	Very High
Speed	High	Medium	High	Low
Data required	Medium	Medium	Low	High
Easiness to use	High	Low	High	Very Low

4. Conclusions and directions for future research

This paper presents a literature review of the methods proposed in literature to forecast CT in SMS. In line with our RO, the contributions of this study are mainly two: (i) an updated taxonomy of the methods and (ii) a comparison between them. These results can be interesting for both researchers and managers. Researchers can benefit from this updated taxonomy to understand which new methods and techniques that belong to the area of AI have been proposed, and where there are interesting avenues for new research. Managers can benefit from this work since they have a "map" (the taxonomy) of the possible methods and, when asked to choose which one to invest on, they have a "compass" (the qualitative comparison) to support their decision-making process.

The analysis of the results shows interesting insights regarding the methods. The most popular type of techniques proposed in the considered article is represented by hybrid methods (which were tested in 47% of the examined papers), followed by ANN (32%), RT (6%) and QT (6%). In all the studies analysed, hybrid methods achieved higher accuracies than the other ones, particularly MLR. Hence, it is likely that scholars will look at more advanced hybrid methods also in the next years. As it was already mentioned in section 3.1.4, BPN are the most popular type of ANN. However, the two most recent works analysed – Tin et al. (2019) and Wang et al. (2018) – both employ LSTM networks, a kind of RNN.

LSTM algorithms were applied to predict traffic flows and have been recently employed also to CT forecasting in SMS, since there are several analogies between the arrival of the cars to a crossroad and the arrival of the lots to a machine. However, as it was explained in the first of the two aforementioned articles, it is still necessary to increase the size of the historical datasets, as well as to increase the number of hidden layers. Moreover, computational times are still very high: for example, since it took 8 hours to complete one full experiment. To tackle this issue, the authors decided to apply Parallel Computing. A similar idea is present in Chen and Wu (2017), where Cloud Computing was employed.

In the future, it could be interesting to apply a fuzzy logic to LSTM algorithms, or to try new hybrid combinations (e.g. FCM-FLSTM) in order to achieve even higher accuracies.

However, since accuracy is not the only driver in CTF, it is also necessary to look for new solutions in terms of speed, robustness, and easiness to use. In this sense, it would be interesting to try a hybrid method like a KM-BPN to a real, large dataset. Interestingly, according to our ranking, accuracy and speed appear to be in trade-off for all the CTF methods. This can represent a pivotal point for a decision maker to orient the choice of the proper method adopt in a specific context. Nevertheless, in line with what discussed in the previous sections, the time required to build and execute the model should be evaluated in light of its magnitude compared to the cycle time of a job and the necessary frequency of update (which is linked with the frequency of the arrival of new orders in the system). On the other hand, there are marginally decreasing improvements in accuracy, and thus, after a certain threshold, the improvement is not substantially appreciable. In light of these considerations, a broader study could be carried out to provide a more elaborate version of the ranking proposed in this paper, for example considering also the different relative importance of the performance dimensions through a system of weights assigned to the different performance. Moreover, the different methods themselves need to be re-assessed, because of the recent advancements of the hardware, which provide a much higher computational power than the past. Finally, some comparisons not deeply investigated in the literature need to be taken into account, such as Queuing Theory or Time Series versus Hybrid methods.

This paper presents some limitations. The first is that we excluded conference proceedings from the search, despite they often present the most recent advancement of the research. Secondly, a more in-depth analysis of the variables considered in each method can be performed, to better confront and contrast the different methods. Future research will be devoted to fill these gaps.

References

Backus, P., Janakiram, M., Mowzoon, S., Runger, C., & Bhargava, A. (2006). Factory cycle-time prediction with a data-mining approach. *IEEE Transactions on Semiconductor Manufacturing*, *19*(2), 252-258.

Chang, P. C., & Liao, T. W. (2006). Combining SOM and fuzzy rule base for flow time prediction in semiconductor manufacturing factory. Applied Soft Computing, 6(2), 198-206.

Chang, P. C., Fan, C. Y., & Wang, Y. W. (2009). Evolving CBR and data segmentation by SOM for flow time prediction in semiconductor manufacturing factory. *Journal of Intelligent Manufacturing*, *20*(4), 421.

Chen, T. (2003). A fuzzy back propagation network for output time prediction in a wafer fab. *Applied Soft Computing*, *2*(3), 211-222.

Chen, T. (2007). An intelligent hybrid system for wafer lot output time prediction. *Advanced Engineering Informatics*, *21*(1), 55-65.

Chen, T. (2007). Incorporating fuzzy c-means and a back-propagation network ensemble to job completion time prediction in a semiconductor fabrication factory. *Fuzzy Sets and Systems*, *158*(19), 2153-2168.

Chen, T. (2008). An intelligent mechanism for lot output time prediction and achievability evaluation in a wafer fab. *Computers & Industrial Engineering*, *54*(1), 77-94.

Chen, T. (2009). A fuzzy-neural knowledge-based system for job completion time prediction and internal due date assignment in a wafer fabrication plant. *International Journal of Systems Science*, *40*(8), 889-902.

Chen, T. (2011). Job cycle time estimation in a wafer fabrication factory with a bi-directional classifying fuzzy-neural approach. *The International Journal of Advanced Manufacturing Technology*, *56*(9-12), 1007-1018.

Chen, T., & Lin, Y. C. (2009). A fuzzy back propagation network ensemble with example classification for lot output time prediction in a wafer fab. *Applied Soft Computing*, *9*(2), 658-666.

Chen, T., & Romanowski, R. (2013). Precise and accurate job cycle time forecasting in a wafer fabrication factory with a fuzzy data mining approach. *Mathematical Problems in Engineering*, *2013*.

Chen, T., & Wu, H. C. (2017). A new cloud computing method for establishing asymmetric cycle time intervals in a wafer fabrication factory. *Journal of Intelligent Manufacturing*, *28*(5), 1095-1107.

Chen, T., Wang, Y. C., & Tsai, H. R. (2009). Lot cycle time prediction in a ramping-up semiconductor manufacturing factory with a SOM–FBPN-ensemble approach with multiple buckets and partial normalization. *The International Journal of Advanced Manufacturing Technology*, *42*(11-12), 1206-1216.

Chen, T., Wu, H. C., & Wang, Y. C. (2009). Fuzzy-neural approaches with example post-classification for estimating job cycle time in a wafer fab. *Applied Soft Computing*, *9*(4), 1225-1231.

Chiu, C., Chang, P. C., & Chiu, N. H. (2003). A case-based expert support system for due-date assignment in a wafer fabrication factory. *Journal of Intelligent Manufacturing*, *14*(3-4), 287-296.

Ciccullo, F., Pero, M., Caridi, M., Gosling, J., & Purvis, L. (2018). Integrating the environmental and social sustainability pillars into the lean and agile supply chain management paradigms: A literature review and future research directions. *Journal of Cleaner Production*, 172, 2336-2350.

Connors, D.P., Feigin, G.E. and Yao, D.D. (1996). A queueing network model for semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, 9(3), 412–427.

Dabbas, Russ M., and J. W. Fowler. 2003. "A new scheduling approach using combined dispatching criteria in wafer fabs." *IEEE Transactions on Semiconductor Manufacturing* 16(3), 501-510.

Gosling, J., & Naim, M. M. (2009). Engineer-to-order supply chain management: A literature review and research agenda. *International journal of production economics*, *122*(2), 741-754.

Hsu, P. L., Hsu, C. I., Chang, P. C., & Chiu, C. (2006). Regression trees approach for flow-time prediction in wafer manufacturing processes using constraint-based genetic algorithm. *International journal of production research*, *44*(24), 5327-5341.

Hung, Y. F., & Chang, C. B. (1999). Using an empirical queueing approach to predict future flow times. *Computers & industrial engineering*, *37*(4), 809-821.

Kuo, C. J., Chien, C. F., & Chen, J. D. (2011). Manufacturing intelligence to exploit the value of production and tool data to reduce cycle time. *IEEE Transactions on Automation Science and Engineering*, *8*(1), 103-111.

Mönch, L., Fowler, J. W., Dauzère-Pérès, S., Mason, S. J., Rose, O., (2011). A survey of problems, solution techniques, and future challenges in scheduling semiconductor manufacturing operations. *Journal of Scheduling*, 14(6), 583-599.

Sha, D. Y., & Hsu, S. Y. (2004). Due-date assignment in wafer fabrication using artificial neural networks. *The International Journal of Advanced Manufacturing Technology*, *23*(9-10), 768-775.

Tin, T. C., Chiew, K. L., Phang, S. C., Sze, S. N., & Tan, P. S. (2019). Incoming Work-In-Progress Prediction in Semiconductor Fabrication Foundry Using Long Short-Term Memory. *Computational intelligence and neuroscience*, *2019*.

Tirkel, I. (2013). Forecasting flow time in semiconductor manufacturing using knowledge discovery in databases. *International Journal of Production Research*, *51*(18), 5536-5548.

Wang, J., & Zhang, J. (2016). Big data analytics for forecasting cycle time in semiconductor wafer fabrication system. *International Journal of Production Research*, *54*(23), 7231-7244.

Wang, J., Yang, J., Zhang, J., Wang, X., & Zhang, W. (2018). Big data driven cycle time parallel prediction for production planning in wafer manufacturing. *Enterprise information systems*, *12*(6), 714-732.