applied sciences

MDPI

Review

A Survey of Feature Set Reduction Approaches for Predictive Analytics Models in the Connected Manufacturing Enterprise

Phillip M. LaCasse [1,*], Wilkistar Otieno [1] and Francisco P. Maturana [2]

[1] Department of Industrial and Manufacturing Engineering, University of Wisconsin–Milwaukee, Milwaukee, WI 53211, USA; otieno@uwm.edu

[2] Rockwell Automation, Inc. Milwaukee, WI 53204, USA; fpmaturana@ra.rockwell.com

* Correspondence: placasse@uwm.edu

Received: 9 January 2019; Accepted: 22 February 2019; Published: 27 February 2019

Abstract: The broad context of this literature review is the connected manufacturing enterprise, characterized by a data environment such that the size, structure and variety of information strain the capability of traditional software and database tools to effectively capture, store, manage and analyze it. This paper surveys and discusses representative examples of existing research into approaches for feature set reduction in the big data environment, focusing on three contexts: general industrial applications; specific industrial applications such as fault detection or fault prediction; and data reduction. The conclusion from this review is that there is room for research into frameworks or approaches to feature filtration and prioritization, specifically with respect to providing quantitative or qualitative information about the individual features in the dataset that can be used to rank features against each other. A byproduct of this gap is a tendency for analysts not to holistically generalize results beyond the specific problem of interest, and, related, for manufacturers to possess only limited knowledge of the relative value of smart manufacturing data collected.

Keywords: connected enterprise; smart manufacturing; big data; machine learning; data reduction; predictive analytics

1. Introduction

In exploring recent advancements in manufacturing and industry, interested practitioners and researchers might find themselves deciphering a series of seemingly related, sometimes interchangeable, but actually distinct terms that mean different things to different parties in different contexts. In some cases, specific terminology might be used in one part of the world whereas another term is employed elsewhere. Consider the following examples, representative but not exhaustive, that are commonly seen today.

One familiar idiom, "smart manufacturing", is a general term for the use of sensors and wireless technologies to capture data in all stages of production or product lifecycle. Examples include vehicle engines collecting and transmitting diagnostic information or optical scanners detecting defects in printed circuits [1,2].

Another common term, "Industrial Internet of Things (IIOT)", initially coined by General Electric (GE) in 2012, refers to a network of industry devices connected by communications technologies for the purposes of monitoring, collection, exchange, analysis and delivery of insights to drive smarter, faster business decisions [3,4]. Examples of consortia targeting the IIOT include the Industrial Internet Consortium (IIC) [5] and the OpenFog Consortium [6].

Industry 4.0 [7,8], China Manufacturing 2025 [9], and Connected Industries [10,11] are specific paradigms of the IIOT applied in the manufacturing context with common fundamental concepts such as smart manufacturing, cyber-physical systems, self-organization, adaptability and corporate social responsibility.

Finally, the Connected Enterprise (CE), a Rockwell Automation term [12] to describe its vision for the future of industrial automation, is an enterprise-wide extension of Industry 4.0 and China Manufacturing 2025. The integration of information technology (IT) and operational technology (OT) enables the interaction between manufacturing process data, people, and the business enterprise to optimize key performance indicators (KPI) at all levels of the organization: factory level, enterprise level and global supply chain level. Figure 1 is a pictorial representation of the CE strategy that enables the seamless convergence of the information and operation functions of an enterprise.

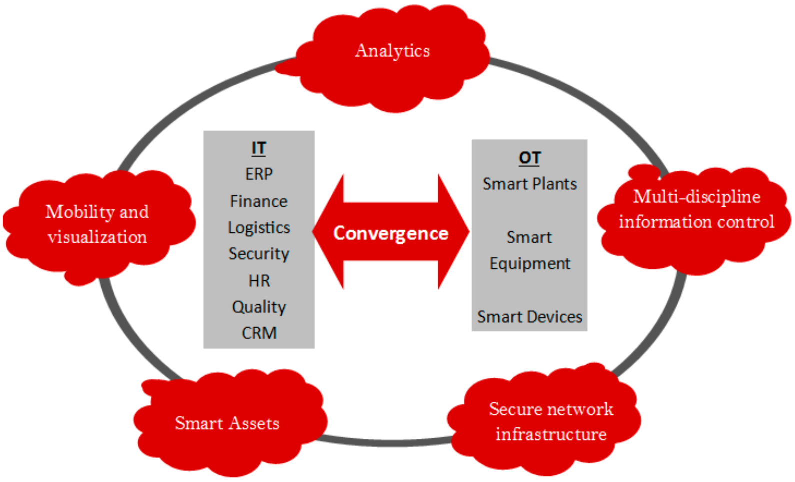

Figure 1. The Connected Enterprise strategy—enabling information technology (IT)—operational technology (OT) convergence.

The Connected Enterprise is one of: [13]:

- enterprise-wide visibility and collaboration;
- interconnected people, equipment, and processes;
- real-time learning of enterprise status;
- organizational agility by means of increased information to make informed, adaptive, proactive decisions.

The purpose of this literature review is to survey and discuss representative examples along the spectrum of existing research into a specific key enabler to the Connected Enterprise in manufacturing: big data. A good working definition of a "big data" environment is one such that the size, structure, or variety of information strains the capability of traditional software or database tools to capture, store, manage, and analyze it [14,15]. Big data is clearly both an enabler to the CE and an obstacle. It is an obstacle in that, by definition, it requires innovation to truly harness; it is an enabler in that it is precisely the availability of vast untapped data that undergirds the enormous potential of the CE. Bollier (2010) explores this duality in big data for The Aspen Institute in [16].

Three related factors provide the motivation for this research. First, the rapid advent of technology to capture and store manufacturing data without the parallel development of corresponding analytical capabilities has resulted in the circumstance by which vast quantities of data are collected but not effectively analyzed or interpreted [1]. Second, the dual nature of big data as both an enabler and an obstacle to the CE perpetuates the state of affairs by which manufacturers do not have a clear picture as to what manufacturing data, of the vast volumes collected, is truly valuable versus what can be discarded. This lack of clarity is due to disjoint or "siloed" data analytics capabilities within the organization [17] and by "where-to-start" paralysis brought on by the sheer volume of data and underdeveloped capability to visualize it [18]. Finally, it has been well established that limitations exist in how and to what extent the human analyst can process complex information [19–21]. The astounding development of the capabilities of automated analytical and artificial intelligence tools prompts interest in steps that can be taken to make them more palatable and digestible to human analysts and decision makers.

The objective of this literature survey is to determine whether, among the extensive body of knowledge on big data in the manufacturing environment, there is room for research into mechanisms or frameworks for feature filtration and prioritization when building applied machine learning models for predictive analytics. The interest is not so much in technology or architecture, which has been explored elsewhere [22–25], but rather in the human factor and how enablers to individual competencies can address the enterprise-level motivations for this research. It is true that certain machine learning algorithms can accommodate high dimensionality in input data, at the risk of potential issues such as overfitting. However, simply incorporating every potential feature into the model because it can algorithmically handle the calculations can shortchange the organization out of potentially useful information about the data at its disposal. The optimal subset problem is NP-Hard, which makes it impractical to iterate through all possible subsets of features to find the best subset for model training. For this reason, there is practical benefit in identifying how analysts and data scientists in manufacturing organizations decide how to select features for model inclusion and if that process is algorithmic and generalizable or if it is ad hoc, tailored to the specific problem of interest.

The remainder of the paper breaks down as follows. Section 2 describes the methodology employed in identifying which articles to review and how to group them. Section 3 contains the literature survey divided into three subsections. Each article receives short commentary in isolation regarding its applicability to the research motivations or objective. At the end of each subsection, a short discussion provides consolidated observations. Section 4 contains discussion with observations spanning the three subsections, and Section 5 provides brief concluding remarks.

2. Methodology

2.1. Data Collection

Articles in this review can be broadly categorized into two groups. The first group consists of featured articles that receive analysis and discussion as pertaining to the motivations and/or objective of this research. The second group consists of background or supporting work that provides context to the introduction, justification to the motivations, or theoretical foundations to techniques or algorithms referenced in the first group.

The process of identifying articles for the first group began with broad queries into databases of scholarly literature using a series of topically relevant keywords. The following keywords were used, typically in pairs but sometimes in groups of three or more: [“big data”], [“smart manufacturing”], [manufacturing], [“machine learning”], [deep learning], [“deep learning”], [“fault detection”], [“fault prediction”], [“fault diagnosis”], [“data reduction”], [“feature selection”], [“feature reduction”], [“instance selection”], and [“instance reduction”]. Quotation marks indicate that the phrase was searched in its entirety. Thus, the keyword [“deep learning”] would not return the phrase “deep neural network learning” but the keyword [deep learning] would.

Academic or scholarly databases searched include ScienceDirect, Institute of Electrical and Electronics Engineers (IEEE), Taylor & Francis, SpringerLink, Google Scholar, and the University of Wisconsin—Milwaukee library system. Time parameters were set for 2008 through 2018.

The keywords employed in database searches were selected to initially catch a wide scope of articles and then converge towards articles focusing more directly on the motivations and objective of the review.

A second means of identifying articles was to survey citations in articles identified in the database searches. For example, if a database search identified a survey paper on the use of machine learning for smart manufacturing, it would be possible that the articles cited therein might pose some relevance.

The intent is not to duplicate work but rather to complement it. Returning to the previous example, a list of articles analyzed from an algorithmic perspective on which machine learning technique was employed could be relevant to this review by seeing how those same papers approach the human dimension of the project.

To identify citations in the second category of articles, the process was ad hoc and tailored to the specific algorithm or technique that warranted additional background. This category made no restrictions to time window because many techniques employed today have their theoretical foundations in decades past. For example, much of the initial, exploratory research into human limitations in processing information took place decades ago.

2.2. Data Analysis

The first layer of analysis consisted of broadly categorizing or organizing the reviewed articles. In keeping with the general search methodology, this resulted in three general groups, selected for their intuitive sense in logically flowing from a broad, high-level search and then converging on the motivations and objective for the review.

- The first category explores big-data models for general industrial applications, specifically those featuring machine learning or deep learning.
- The second category focuses specifically on big data analyses and frameworks as applied to scenarios specific to smart manufacturing. Two subtopics emerged in the search results: fault detection and fault prediction.
- The third category addresses data reduction tools and techniques.

The three categories listed above came about partly by design and partly post hoc. From the beginning, the question of interest was data reduction, specifically feature filtration and prioritization. Upon conducting a high-level analysis of articles captured by queries described in Section 2.1, it became clear that it would be appropriate to organize by papers explicitly focused on data reduction and those not. Clearly, a paper that is explicitly on the topic of data reduction will cover the subject. However, this review is also interested in how articles approach the topic of data reduction as a step contained within some problem of interest, when the paper is not explicitly about data reduction. This would have resulted in two categories. It subsequently became clear upon examination that, of the papers not explicitly focused on data reduction, they could be subdivided into those focused on a specific manufacturing application and those focused on general applications independent of a specific problem type. This yielded the three categories that ultimately form the organization of Section 3.

The second layer of analysis consisted of identifying which articles merit discussion and how to organize that discussion.

The predominant theme for analyzing articles in the first category, general industrial applications, was the degree to which the article focused on enterprise capabilities that enable organizational competencies versus approaches or methodologies that relate to human competencies. The first two motivations for this research are predominantly organizational competencies that are developed by a combination of high-level, enterprise capabilities and low-level, individual competencies. Of interest to this review was whether the reviewed articles gave treatment to the research motivations and, if so, whether that treatment focused on the organizational or the individual competencies.

The focus for analyzing articles in the second category, specific manufacturing applications, was the extent to which data reduction was explicitly performed and, if so, the extent to which that reduction step received treatment in terms of analysis or generalizability. The working hypothesis was that most research would be focused on a specific application or problem of interest, with the input data treated in secondary fashion, being a means to some end and not as potentially an end unto itself. The reasoning behind the working hypothesis is that practitioners and researchers alike have priorities of work; solving the problem of interest is typically Priority #1. Time-constrained efforts to complete the task at hand can sometimes cause both researchers and practitioners alike to miss valuable nuggets of insight that could provide useful in subsequent future work.

The focus for the third category was the context for the data reduction and the type of data reduction performed. If the context was outside of the manufacturing realm, the question was if it would be possible to extend the technique to manufacturing contexts. If already contextualized within manufacturing, the question was how generalizable it might be to other contexts or if the technique was unique to the specific scenario or case study.

Within each section, individual articles receive commentary in isolation. Each section concludes with observations and discussion on themes contained in more than one paper therein. Finally, Section 4 provides consolidated observations and discussion for the entire set of reviewed literature.

3. Literature Survey

3.1. Big Data Approaches for General Industrial Applications

Existing research into big data utilization for general industrial applications may be broadly generalized to contain valuable work and insight into the state of technology, current challenges, and methodologies or high-level frameworks for big-data analytical projects. The following section contains examples that, while not intended to be exhaustive, are representative of the body of literature on the subject. These examples are reviewed with specific interest in how they treat the research motivations from the human versus the architectural or technological dimension.

Wuest et al. (2016) present an overview of machine learning in manufacturing, focusing specifically on advantages, challenges, and applications [26]. Of particular interest is a summary of several recent studies ([27–30]) on the key challenges currently faced by the larger global manufacturing industry, with agreement on the following key challenges:

- Adoption of advanced manufacturing technologies
- Growing importance of manufacturing of high value-added products
- Utilizing advanced knowledge, information management, and AI systems
- Sustainable manufacturing (processes) and products
- Agile and flexible enterprise capabilities and supply chains
- Innovation in products, services, and processes
- Close collaboration between industry and research to adopt new technologies
- New manufacturing paradigms.

It is interesting to observe, in addition to what is listed, what is not listed. Specifically, these recent studies did not identify data dimensionality as a key challenge. In other words, while there is recognition that voluminous manufacturing data is collected, there is not universal agreement that this is a problem that needs to be addressed on the front end [26]; rather, employment of various machine learning techniques is proposed as a means to deal with it [31], with methods towards this end dating as far back as the 1970s [32].

However, employment of machine learning algorithms to deal with the problem of high dimensionality can lead the analyst directly into one of the main challenges associated with machine learning that the paper identifies, which is that interpretation of results can be difficult. Especially when the model is intended to support real-time monitoring of parameters with respect to proximity to some threshold, the practical usefulness of the model is diminished when large numbers of irrelevant or redundant features are input into the model simply because the machine learning algorithm can accommodate them.

Alpaydin (2014) provides a comprehensive overview of machine learning, with specific techniques that apply to each of the needs described above [33]. It is pointed out, however, that existing applications of machine learning tend to narrowly focus on the problem at hand or on a specific process [34] and not holistically on the manufacturing enterprise or on generalizing the results to other processes. This observation is noteworthy, as it relates tangentially to the motivation for this literature review. One reason for the willingness to select machine learning algorithms that can handle high dimensionality may be a 'prisoner-of-the-moment' mentality. Analysts and data scientists perform real-world analyses to solve real-world problems, usually on a deadline imposed beyond their control. That deadline may be imposed by supervisors or it may be a function of outside constraints. Circumstances may not afford the luxury to step back, after completing the initial project, and thoroughly comb through the data to draw secondary conclusions about the nature of the input data. Rather, it is on to the next problem.

Wang et al. (2018) unpack the benefits and applications of deep learning for smart manufacturing, identifying benefits that include new visibility into operations for decision-makers and the availability of real-time performance measures and costs [35]. The authors provide, in addition to this practical information, a useful discussion on deep learning as a big-data analytical tool. In particular, they compare deep learning with traditional machine learning and offer three key distinctions between the two. Those distinctions are summarized in Table 1 [35].

Table 1. Distinction between traditional machine learning and deep learning.

Technique	Feature Learning	Model Construction	Model Training
Traditional Machine Learning	Features are identified, engineered and extracted manually through domain expert knowledge.	Models typically have shallow structures (few hidden layers) and are data-driven using selected features.	Modules are trained step by step.
Deep Learning	Features are learned by transforming the data into abstract representations.	Models are end-to-end, high hierarchies with nonlinear combinations of numerous hidden layers.	Model parameters are trained jointly.

Note the distinction in feature learning. Deep learning models do not explicitly engineer and extract features. Rather, they are learned abstractly. This is both an advantage and a tradeoff. The blessing is that model performance is typically superior. The tradeoff is in the transparency, traceability, and front-end verifiability of results.

The authors make an interesting observation, in that deep learning has shown itself to be most effective when it is applied to limited types of data and well-defined tasks [35]. This is notable in that conventional wisdom sometimes holds more data is better. Reducing the large data set to the most relevant subset of predictors may actually improve performance. This speaks directly to the motivation for this review and demonstrates the importance of the question. Not only does the capability to reduce a feature set to only the most relevant features enable an organization to build and increase institutional knowledge about the data at its disposal, but it also may lead to superior model performance.

Closely related, Tao et al. (2018) provide a comprehensive look at data-driven smart manufacturing, providing a historical perspective on the evolution of manufacturing data, a development perspective on the lifecycle of big manufacturing data, and a framework envisioning the future of data in manufacturing [2].

An observation is that Tao et al. also identify a gap and promising future research direction that aligns indirectly with the focus of this literature review: edge computing. Edge computing is, architecturally, an option for whittling down the volumes of production data into the core pieces that are truly meaningful and align with the key performance indicators (KPIs) of interest. Edge computing allows data to be analyzed at the "edge" of a network before being sent to a data center or cloud [36]. A related term, fog computing, was introduced by Cisco systems in 2014 and extends the cloud to be closer to devices that produce and act on IIOT data [37]. The distinction between the two concepts, as well as other emerging paradigms such as mobile edge computing (MEC) and mobile cloud computing (MCC) are not fully mature and are subject to overlap [38]. The commonality is that they represent means for an organization to operationalize the individual competencies that are the focus of this review.

A final framework for general industrial application of big data is presented by Flath and Stein (2017), specifically in the form of a data science "toolbox" for manufacturing prediction tasks. The objective is to bridge the gap between machine learning research and practical needs [39]. Feature engineering is identified as an important step that must take place prior to deriving useful patterns from the input data, and a case study employs Kullback–Leibler divergence to reduce 968 numeric features to 150 and 2140 categorical features to 27.

The preceding literature, summarized in Table 2 below, shows high-level analysis of trends and challenges. It also provides examples of methodologies and frameworks for applied big data analytics in manufacturing.

Table 2. Summary—big data for general industrial applications.

Author(s)	Focus
Wuest et al. [26]	Key challenges for global manufacturing industry
Alpaydin [33]	Machine learning overview
Wang et al. [35]	Deep learning for smart manufacturing
Tao et al. [2]	Data-driven smart manufacturing
Flath and Stein [39]	Data science "toolbox" for industrial analytics

A first observation is that there is not uniform agreement with regard to the question of dimensionality. At one extreme, the question is treated as a non-issue, to be handled by the machine learning algorithm selected. Other articles addressed the question at a high-level as important but always within the context of the larger problem-solving approach and not to the level of detail that would be useful to the data scientist.

A second observation is that the approaches for predictive analytics in this section are geared less towards the detailed steps that an analyst might perform and more towards the infrastructure, architecture, and general data landscape that an organization should possess in order to have the capability to perform applied predictive analytics projects. This is not entirely unexpected, as the articles in this section are selected specifically for their high-level, broad outlook. The expectation is that articles in Sections 3.2 and 3.3 will provide greater detail on the subject because articles reviewed in those sections focus more precisely on contexts that align better to activities at the level of the analyst or data scientist.

A third observation is gap identified by more than one researcher, which is the lack of holistic generalization of results beyond the specific, local problem under examination. This is related to manufacturers' limited knowledge of the relative utility or value contained among the different elements of the vast volumes of data that they collect in a somewhat mutually-reinforcing way. A lack of knowledge regarding the data landscape makes it difficult to generalize a dataset's utility from one application to the next. On the same token, not taking incremental steps to analyze projects after the fact for relevance and generalizability to other contexts perpetuates the deficiency in institutional knowledge.

3.2. Big Data Approaches for Specific Manufacturing Applications

This section moves from the higher level of general industrial or manufacturing applications to approaches geared towards specific smart manufacturing applications. The following literature instances fall into one of two subcategories: fault detection and fault prediction. Fault detection and fault prediction are important areas of interest, and it is not surprising that predictive analytics projects gravitate to those topics. Predictive analytics in any the context will naturally gravitate to the dominant interests or challenges facing decision makers in that context, and, for manufacturers, key performance indicators (KPIs) associated with cost, quality, and time are negatively influenced by faults in machinery or output. Most manufacturing processes involve some form of creation or assembly at a given stage followed by some manner of inspection or validation before moving on to the next stage. Components are assembled into some final product, which itself undergoes functional testing prior to distribution to the customer. Machine downtime for unscheduled maintenance will negatively impact cycle time and, by extension, cost. Undetected malfunctions or nonconformities in machinery can lead to defective products escaping from one stage of manufacture to the next. There is an ever-present need to reduce defective products, which creates a natural partnership between smart manufacturing and predictive analytics. It is therefore unsurprising that much of the literature in predictive analytics in the manufacturing context will be applied to case studies in either fault detection or fault prediction.

It will be observed that different publications employ different frameworks, techniques, models, or methodologies to address specific manufacturing applications, often addressing specialized subproblems or challenges. The focus in the ensuing section is how, from the human data scientist perspective, these analyses approach the challenge posed by big data. Is the big data challenge one of an excessive number of diverse features that may contain hidden predictive potential? Is the challenge one of data volume, with exceedingly large numbers of records produced? Neither? Both? Additionally, this review will analyze the ensuing articles with an eye towards knowledge management, or the extent to which there is opportunity to generalize beyond the specific problem of interest.

3.2.1. Fault Detection

In [40], a MapReduce framework is proposed and applied to the fault diagnosis problem in cloud-based manufacturing under the circumstance of a heavily unbalanced dataset. An unbalanced dataset is one in which a large number of examples but another class is represented by comparatively far fewer [41,42]. In terms of features for use in model training, each record of input data contains 27 independent variables and one fault type. There is no explicit discussion of reducing the 27 input variables to a smaller subset or what steps might be taken to do so for a scenario with higher dimensionality.

A hybridized CloudView framework is proposed in [43] for analyzing large quantities of machine maintenance data in a cloud computing environment. The hybridized framework contrasts with a global or offline approach [44] and a local or online approach [45], providing the advantage of being able to analyze sensor data in real-time while also predicting faults in machines using global information on previous faults from a large number of machines [43]. Feature selection is discussed at a high level, but the illustrative case study employs only three data inputs. The purpose of the case study is simply to illustrate the case-based reasoning applied and not apparently to address a specific situation.

In [46], Tamilselvan and Wang employ deep belief networks (DBN) for health state classification of manufacturing machines, with IIOT sensor data employed for model inputs. Specifically, signal data from seven different signals out of a possible 21 were selected for model training. Selection of which signals to include for model training was made based on literature and not on a specific methodological approach.

Deep belief networks are compared favorably to support vector machines (SVM), back-propagation neural networks (BNN), Mahalanobis distance (MD), and self-organizing maps (SOM) [46]. The deep belief network structure consists of a data layer, a network layer, and some number of hidden layers in between. This particular framework structures its hidden layers as a stacked network of restricted Boltzmann machines (RBMs) [47], with the hidden layer of the n^{th} RBM as the data layer of the (n+1)th RBM.

A similar machine learning methodology is employed by Jia et al. (2016) for fault characterization of a rotating machinery in an environment characterized by massive data using deep neural networks (DNNs) [48]. A DNN is similar to the DBN, except that the layers are not constrained to be RBM. For an extensive overview of deep learning in neural networks, see [49]. In a case study in fault diagnosis of rolling element bearings, a total of 2400 features are extracted from 200 signals using fast Fourier transform (FFT); no explicit reduction step is performed or discussed. Rather, the full dataset is input into the DNN.

The DNN model achieves impressive results when compared with a back-propagation neural network (BPNN), with correct classification rates over 99% compared to 65–80% for the BPNN [48]. This indicates that the specific algorithm employed can have a non-trivial impact on the results, depending on the problem under study.

A framework for fault signal identification is proposed by Banerjee et al. (2010) in [50] using short term Fourier transforms (STFT) to separate the signal and SVM to classify it, and Banerjee and Das (2012) extend the approach in [51]. An explicit discussion on data preparation or feature filtration is absent due to the manageable feature set used for model training. However, the approach to extract features from signal data can lead to an excessive number of potential features, making such a step value-added.

Note also that this framework is a hybrid of several techniques, taking sensor data into the SVM after it has already been processed by signal processing and the time-based model. This is in contrast to frameworks relying exclusively on SVM [52,53] or exclusively on time series analysis [54].

Probabilistic frameworks for fault diagnosis grounded in Bayesian networks (BN) and the more generalized Dempster–Shafer theory (DST) are examined in [55] and [56], respectively. For background and additional information on DST, see [57]. The challenge explored by Xiong et al. (2016) in [56] is that of conflicting evidence, with the observation that, in practice, sensors are often disturbed by various factors. This can result in a conflict in the obtained evidence, specifically in a discrepancy between the observed results and the results obtained by fusion through Dempster's combination rule. This challenge reveals the need to reprocess the evidence using some framework or methodology prior to fusing it. Xiong et al. (2016) propose to do so with an information fusion fault diagnosis method based on the static discounting factor, and a combination of K-nearest neighbors (KNN) and dimensionless indicators [56].

Just as in Jia et al. (2016), Xiong et al. (2016)'s method is applied to fault diagnosis among rotating machinery in a large-scale petrochemical enterprise.

Khakifirooz et al. (2017) employ Bayesian inference to mine semiconductor manufacturing data for the purposes of detecting underperforming tool-chamber at a given production time. The authors use Cohen's kappa coefficient to eliminate the influence of extraneous variables [58].

The tool-chamber problem examined in [58] is relevant to this review in that it employs a large number of binary input variables in its model, one for each tool and each step, equal to 1 if the tool-chamber feature was used in a step and equal to 0 if not. The feature filtration approach employed is a two-fold application of Cohen's kappa coefficient, once for pairwise comparison of the features against each other and once for features against the target. Features exhibiting high agreement with each other are wrapped with peers into a group; feature exhibiting low agreement with the target are removed from the model, with 0.20 as the threshold for removal.

This method is appropriate when features and the target are both binary; a limitation is the method is not suitable for data in other forms. This required the target to be transformed from a continuous yield percentage to a categorical classification. A second possible limitation is that each variable is tested independently of the others, with no consideration for interaction. It is logically possible that a feature could have a poor Cohen's kappa coefficient but could interact with other features to produce an overall better model. An advantage of the approach, though not specifically discussed in the article, is that Cohen's kappa coefficient scores for each feature may be preserved from one analysis to the next and analyzed to see if they harbor latent relationships that might point to root causes of inadequate tool-chamber and not simply forecast it.

The final framework for fault detection that this literature review will explore is a cyber-physical system (CPS) architecture proposed by Lee (2017) for fault detection and classification (FDC) in manufacturing processes for vehicle high intensity discharge (HID) headlight and cable modules [59]. For additional background and exploration of CPS, see [60–63]. Although much of the article is devoted to material outside the scope of this review, such as network and database architecture, the manufacturing process explored is notable because it involves multiple subprocesses, some of which are performed in-house and some of which are outsourced to external parties. Furthermore, although there is a small set of main defects that may be observed (shorted cable, cable damage, insufficient soldering, and bad marking), those faults are not directly traceable to a single subprocess. Rather, any number of different subprocesses may result in any fault type. The impact, when performing fault detection and classification, is that the cause-effect relationships and the backwards tracing of faults to diagnoses must take place beforehand.

The input data for the case study consists of eight signals, three from torque sensors and five from proximity sensors, and three learning models are explored: support vector regression (SVR), radial bias function (RBF), and deep belief learning-based deep learning (DBL-DL). In the SVR and RBM models, no additional step in data filtration or feature extraction is performed; in the DBL-DL model, features are extracted in the form of two hidden layers. Unsurprisingly, the DBL-DL model outperforms the other two, with a classification error rate of 7% as compared to 8% for SVR and 9% for RBM [59].

3.2.2. Fault Prediction

In [64], Wan et al. (2017) present a manufacturing big data approach to the active preventive maintenance problem, which includes a proposed system architecture, analysis of data collection methods, and cloud-level data processing. The paper mainly focuses on data processing in the cloud, with pseudocode provided for a real-time processing algorithm. Two types of active maintenance are proposed as necessary: a real-time component to facilitate immediate responses to alarms and an offline component to analyze historic data to predict failures of equipment, workshops or factories.

Of interest to this review is to note that the aforementioned approach is in the context of an organization's ability to perform active preventive maintenance and not in the context of how a data scientist goes about performing his or her analysis. For example, 'data collection' in the context that Wan et al. describe refers to the required service-oriented architecture to integrate data from diverse sources. To the data scientist, 'data collection' is the employment of that architecture in identifying and obtaining specific data elements for model inclusion.

Munirathinam and Ramadoss (2014) apply big data predictive analytics to proactive semiconductor production equipment maintenance. Beginning with a review of maintenance strategies, the researchers present advantages and disadvantages for each of four different maintenance strategies: run to failure (R2F), preventive, predictive, and condition-based. Following this background, an approach for predictive maintenance is presented as follows [65]:

- Collect raw FDC, equipment tracking (ET), and metrology data
- Perform data reduction using a combination of principal component analysis (PCA) and subject matter expertise. This step, in the semiconductor case study, reduces the set of possible parameters from over 1000 to precisely 16
- Train model
- Display output to dashboard with a Maintenance/No Maintenance status

Two immediate observations are apparent when considering the data reduction step employed in this model. First, the use of PCA is effective but it carries with it the loss of interpretability after the fact. This limits the options associated with the dashboards created for visualization of model results. If there were an alternative to PCA that retains interpretability, it may be possible to identify specific thresholds in the input data that are triggers for required maintenance and then track proximity to those thresholds in a dashboard. A second observation is that PCA requires linearity among the parameters because it relies on Pearson correlation coefficients. It also assumes that a feature's contribution to variance relates directly to its predictive power [66]. It is not clear that this is always an appropriate assumption.

Ji and Wang (2017) present a big-data analytics-based fault prediction approach for shop floor scheduling. This application of the big data problem focuses less on the availability of machining resources and more on the problem of potential errors after scheduling [67]. Specifically, it is observed that task scheduling using traditional techniques considers currently available equipment, with time and cost saving as the main objectives. Missing from consideration is the condition prediction of the machines and their states. In other words, scheduling is made absent of any information on the expected condition of the machines during the production process. In the proposed framework, tasks are represented by a set of data attributes, which are then compared to fault patterns mined through big data analytics. This information is then used to assign a risk category to tasks based on generated probabilities. The model provides the opportunity for prediction of potential machine-related faults such as machine error, machine fault, or maintenance states based on scheduling patterns. This knowledge can lead to better machine utilization.

It should be noted that this particular framework, while creative, was not tested on actual data but rather on hypothetical datasets due to data proprietorship policy [67], hence providing clear opportunities for future research.

Neural networks are applied to recognize lubrication defects in a cold forging process, [68] predict ductile cast iron quality [69], optimize micro-milling parameters [70], predict flow behavior of aluminum alloys during hot compression [71], and predict dimensional error in precision machining [72]. Finally, a process approach is taken to improve reliability of high speed mould machining [73].

It was seen in the preceding models featuring NN that data reduction plays a role of minimal importance because the neural network accomplishes feature creation and selection in the hidden layers. In [68], a total of 20 features are selected for model input with no explicit data reduction step. Nor was any reduction step performed in [69], where the dataset was relatively small, consisting of only 700 instances of 14 independent variables in the training set. In [70] and [71], only three features are input into the artificial neural networks (ANN). In [72], an extension of a simulation and process planning approach in [73,74], the number of input variables is five.

Finally, quality and efficiency in freeform surface machining are driven by three primary issues: tool path, tool orientation, and tool geometry [75]. A feature-based approach to machining complex freeform surfaces in the cloud manufacturing environment yields the capability for adaptive, event-driven process adjustments to minimize surface finish errors [76].

An observation across the set of articles reviewed in Section 3.2 is that a specific data reduction step is rarely utilized, either because the feature set was small to begin with or because the machine learning technique could accommodate. The exceptions used either statistical measures (Cohen's kappa) or PCA to reduce the feature set. The article using the former technique did not report how many features the case study began with and how many were ultimately used for model training. It is, therefore, not clear the extent to which the technique is useful. In the case of PCA with subject matter expertise, a feature set of 1000 reduced to 16. Additional discussion and possible extension will be included in Section 4.

A second observation is that as in Section 3.1, variation exists in the frame of reference for which different articles approach the topic of predictive analytics. Some articles focus on the organizational capability to perform predictive analytics. These incorporate robust discussion on technology-centric elements such as architecture for data capture, storage, and extraction or at which levels different analyses may be performed (cloud, edge, real-time, offline, etc.). These typically featured commercially available technologies such as Hadoop or MapReduce and address some of the prerequisites for building organizational competencies in this area. Other articles, on the other hand, employed the term 'framework' to refer to a problem-solving approach or methodology, a sequence of actions to be performed by the analyst or data scientist. These articles more directly align with the objective of this literature review, but it is important to distinguish between the two perspectives as each are important. Indeed, the organizational capability for data capture, storage, and migration must necessarily precede any in-house capability to analyze smart manufacturing data or use it to train a machine learning model.

Table 3 provides a summary of the forgoing studies that approach big data analytics applied to specific manufacturing use cases. The table summarizes whether the paper focuses on organizational capabilities, methodological approaches for the analyst, case studies, or some combination. For case studies, the machine learning algorithm is listed.

Table 3. Big data approaches for specific industrial applications.

Authors(s)	Focus	Explicit Data Reduction Step
Kumar et al. [40]	Enterprise-level architecture; methodology to address class imbalance	No
Bahga and Madiseti [43]	Enterprise-level architecture	No
Tamilselvan and Wang [46]	Case study: Machine health states—DBN	No
Jia et al. [48]	Case study: Fault characterization—DNN	No
Banerjee et al. [50]	Case study: Fault signal identification—SVM	Discussed, not implemented
Xiong et al. [56]	Methodology: Information fusion to reconcile conflicting evidence in fault detection	No
Khakifirooz et al. [58]	Case study: Yield enhancement–Bayesian inference	Yes
Lee [59]	Enterprise-level architecture; Case study: Fault detection and classification—SVR, RBF, DBL-DL	No
Wan et al. [64]	Enterprise-level architecture; Methodology: Real-time and offline components; Case study: Fault prediction—Neural Network	No
Munirathinam & Ramadoss [65]	Enterprise-level architecture	Yes
Ji and Wang [67]	Enterprise-level architecture; Simulated proof of concept case study: Fault prediction for shop floor scheduling	No
Rolfe et al. [68]	Case study: Lubrication defects in cold forging process—NN	No
Perzyk and Kochanski [69]	Ductile cast iron quality—NN	No
Kilickap et al. [70]	Micro-milling parameter optimization—NN	No
Changqing et al. [71]	Alloy flow behavior-NN	No
Arnaiz-Gonzalez et al. [72]	Dimensional error in precision machining—NN	No
de Lacalle et al. [73]	High speed machining of moulds	No
Liu and Li [76]	Manufacturing freeform surfaces	No

3.3. Frameworks for Data Reduction

The third and final category of literature that this review will examine focuses on techniques or approaches specifically for data reduction, which includes feature reduction/selection and instance reduction/selection. There exists a substantial body of influential data preprocessing algorithms for missing values imputation, noise filtering, dimensionality reduction, instance reduction, and treatment of data for imbalanced processing [77]. Specific algorithms for feature selection include Las Vegas Filter/Wrapper [78], Mutual Information Feature Selection [79], Relief [80], and Minimum Redundancy Maximum Relevance (mRMR) [81]. Specific algorithms for instance reduction include condensed nearest neighbor (CNN) [82], edited nearest neighbor (ENN) [83], decremental reduction by ordered projections (DROP) [84], and iterative case filtering (ICF) [85].

The interest in the ensuing articles reviewed in this section is in their suitability for application to the CE. To this end, the domain in which the articles implement any applied case studies is also examined. It will be observed that the reviewed articles contain tasks that fall within Step 3 of the data source selection methodology outlined in [86] and broadly fall into one of three categories: sampling reduction, feature reduction, or instance reduction. Sampling reduction applies to contexts such as optical inspection or reengineering, where there is a need to obtain information for an entire component or surface. If that information may be obtained using fewer samples, then benefits in cost or efficiency follow. Instance reduction applies to contexts in which large numbers of data points are collected for a relatively smaller set of attributes or features. Feature selection is the process of reducing the number of attributes or columns to be input into a machine learning model for training.

Habib ur Rehman, et al. (2016) propose an enterprise-level data reduction framework for value creation in sustainable enterprises, which, while not contextualized to manufacturing, is easily extendable to this domain. The framework considers a traditional five-layer architecture for big data systems and adds three data reduction layers [87].

The first layer for local data reduction is intended for use in mobile devices to collect, preprocess, analyze, and store knowledge patterns. This physical layer can easily be conceptually translated to the CE. The second layer, for collaborative data reduction, is situated prior to the cloud level, with edge computing servers executing analytics to identify knowledge patterns. Note that "edge computing" may be referred to as "fog computing" in some cases [88]. This step will exist in varying degrees in the CE depending on the maturity of the process or organization. In the context of user Internet of Things (IoT) mobile data, as initially presented in the paper, there exists a body of data that must automatically be discarded in accordance with external constrains such as privacy laws. This brings a practical purpose to this initial filtration layer. In smart manufacturing, the physical layer represents IIOT machine or production data, all of which might theoretically harbor some purpose. It may not be prudent to automatically discard chunks of data until it has been definitively determined that there is little risk in doing so. Finally, a layer for remote data reduction is added to aggregate the knowledge patterns from edge servers that are then distributed to cloud data centers for data applications to access and further analyze [87].

It should be noted that this framework is at the institutional level and not at the level of the data scientist. The data reduction layers are presented as automated processes applied to the raw source data and not dependent on a specific project or problem of interest.

At the data scientist level, a second point-based data reduction scenario is presented in [89], in which Ma and Cripps (2011) develop a data reduction algorithm for 3D surface points for use in reverse engineering. In reverse engineering, data is captured from an existing surface on the order of millions of scanned points. There are challenges associated with volume of data, and there are challenges in the form of increased error associated with removing data. The data reduction algorithm is based on Hausdorff distance and works by first collecting a set of 3D point data from a surface using an optical device such as a laser scanner, iterating through the set of points, and determining if a point can be removed without causing the local estimation of surface characteristics to fall out of tolerance. This is done by comparing shape pre- and post-removal. The procedure is tested on an idealized aircraft wing but is extendable to any manner of reverse engineering that employs 3D measurement data. It is possible that this could also be extended to inspection-type applications, but the challenge is that the end-state number of required data points will be dependent on the nature of the surface. Additionally, it is not certain that Hausdorff distance would be the appropriate metric for other contexts such as automated optical inspection.

Considering data reduction with respect to the set of features to be used for model training, Jeong et al. (2016) propose a feature selection approach based on simulated annealing and apply it to a case study for detecting denial of service attacks [90]. This approach is similar to [91], which uses the same data set but a different local search algorithm.

The model starts with a randomly-generated set of features to include, trains a model on that set of features, and tests it by way of some pre-designated machine learning technique. The case study is a classification problem, and so examples used include SVM, multi-layer perceptron (MLP), and naïve Bayes classifier (NBC). After obtaining a solution and measure of performance using some cost function, neighborhood solutions are obtained and tested. Superior solutions are retained, and inferior solutions are either discarded or retained based on a probability calculation. This ability to retain an inferior solution allows the simulated annealing algorithm to "jump" out of a local extrema [92]. The intrusion detection case study employed 41 factors, which reduced to 14, 16, and 19 factors when using MLP, SVM, and NBC respectively. A limitation to this approach is that it requires model training at every iteration of the simulated annealing. This may limit the options for which machine learning technique to select; preference should be given to algorithms that quickly converge. Again, for only 41 factors, this is less of an issue. If there are hundreds or thousands, then this approach may be impractical.

Lalehpour, Berry, and Barari (2017) propose an approach for data reduction for coordinate measurement of planar surfaces for the purposes of reducing the number of samples required to adequately validate that a part has been build according to design specifications [93]. The larger context for this approach is manufacturing, but the applicability is narrowly scoped to an inspection station along an assembly line. Thus, this approach could be used in programming firmware for an optical inspection machine so that it can diagnose defective components as efficiently as possible. However, it would not be useful in performing root cause analysis to find the source of the defects or predict future occurrences.

Ul Haq, Wang, and Djurdjanovic (2016) develop a set of ten features that may be constructed from streaming signal data from semiconductor fabrication equipment. Technological developments allow the collection of inline data at ever increasing sampling rates. This has resulted in two effects, the first being an increase in the amount of data required to store, and the second being the ability to discern features that were previously not discernible [94]. Specifically, high sampling rates allow information to be gleaned from transient periods between steady signal states. This enables the extraction of features from the signal that could not be calculated with lower sampling rates.

The approach can be extended to any signal-style continuous data source from which samples are taken, although the implication is that the lower the sampling rate, the less likely that these new features will provide value. These constructed features are applied to case studies of tool and chamber matching and defect level prediction. A reasonable extension might be to apply the approach to machine diagnostic information for active preventive maintenance.

From a feature selection or dimensionality perspective, which is of most interest to this review, the ten features are calculated every time the signal transitions from one steady state to another. For relatively static signals, this will result in a manageable feature set; for more dynamic signals or for large time windows, the number of calculated features may become prohibitively large. This could be alleviated by adding an additional layer of features that employ various means to aggregate the values of the ten calculated features over the entire span of time.

Continuing on the topic of feature selection, Christ, Kempa-Liehr, and Feindt (2016) propose an algorithm for time series feature extraction, TSFRESH, that not only generates features but also employs a feature importance filter to screen out irrelevant features [95]. This framework, illustrated in Figure 2, begins by extracting up to 794 predefined features from time series data. Subsequently, the vector representing each individual feature is independently tested for significance against the target. This produces a vector of p-values with the same cardinality as the number of features. Finally, the vector of p-values is evaluated to decide which features to keep. The method for evaluating the vector of p-values is to control the false discovery rate (FDR) using the Benjamini-Hochberg procedure [96].

A case study using data from the UCI Machine Learning Repository [97] reduced an initial set of 4764 features to 623 [98].

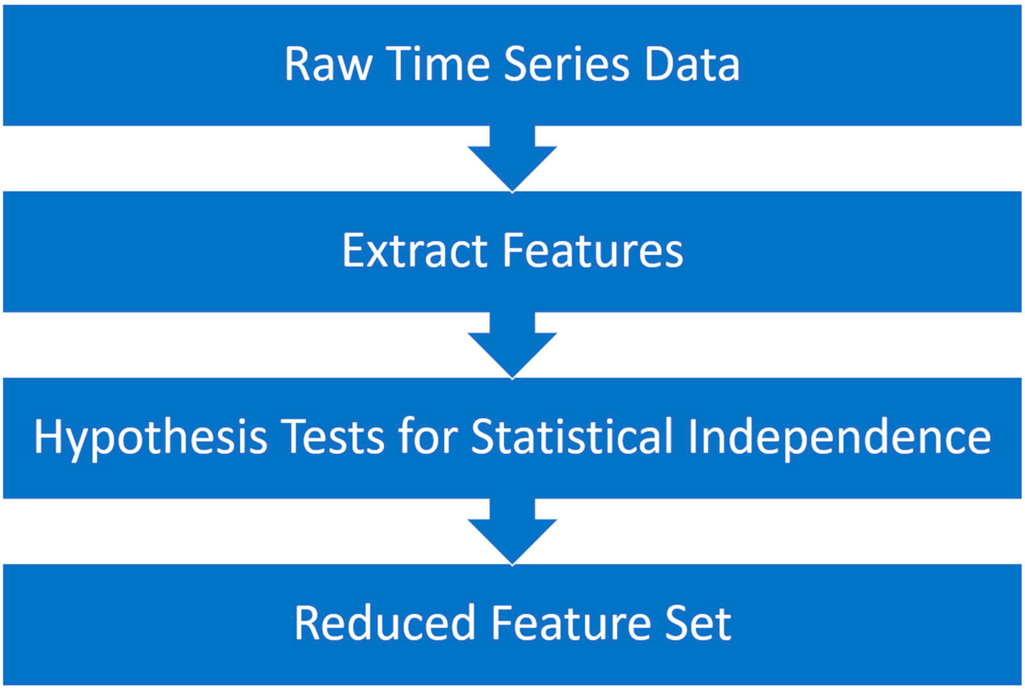

Figure 2. High-level TSFRESH process

For instance reduction, Wang et al. (2016) employ a framework based on two clustering algorithms, affinity propagation (AP) and k-nearest neighbor (k-NN), to extract "exemplars", or representations of some number of actual data points [99]. A clustering algorithm is employed to cluster the data instances into similar groups; an exemplar is then defined to represent the group. The context is in network security, specifically anomaly detection. The idea is that records for http traffic and network data under normal circumstances can be grouped or aggregated into representations of those conditions, which can produce cost savings in data storage. The technique is potentially extendable to other areas of manufacturing, although for mature processes there may not be the desire to perform aggregation of records because the "easy" relationships have already been discovered. Rather, a large number of records may be necessary to identify hidden structures or correlations in subgroups that might otherwise, in smaller sample sizes, be considered outliers [100,101].

A second instance reduction Nikolaidis, Goulermas, and Wu (2010) develop an instance reduction framework that draws a distinction between instances close to class boundaries and instances farther away from class boundaries [102]. The reasoning is that instances farther away from class boundaries are less critical to the classification process and are therefore more "expendable" from an instance reduction standpoint. The four-step framework first uses a filtering component such as ENN to smooth class boundaries and then classifies instances as "border" and "non-border". Following a pruning step for the border instances, the non-border instances are clustered using mean shift clustering (MSC).

As previously indicated, the reviewed articles from the Section 3.3, summarized in Table 4, cover reductions in the number of samples required to obtain a satisfactory result, techniques to reduce the number of instances or records, and techniques to reduce the number of features or attributes.

Of greatest interest to this review is the second category, feature selection, and two approaches seen in this section merit further discussion in relation to each other. The first approach, the TSFRESH approach, generates a list of up to 794 features from a single time series and, using statistical independence as the test, reduces the feature set by eliminating the features that do not exhibit a significant statistical dependence with the response. Using this approach, a model with N time series inputs would have 794 N features extracted by TSFRESH. Even if TSFRESH then filters out 50% of the features, there still could remain many hundreds of features in the model. This could be an excessive number of features that strains the capacity of the analyst to truly grasp what is going on or pinpoint the critical relationship(s) of interest. Extending the approach to include subsequent filter(s) could be a step in remedying this challenge.

Table 4. Frameworks for data reduction.

Author(s)	Focus
Habib ur Rehman, et al. [87]	High level/Institutional framework
Jeong et al. [90]	Feature selection meta-heuristic (simulated annealing)
Lalehpour, Berry, and Barari [93]	Sample reduction
Ma and Cripps [89]	Shape preservation with data reduction for 3D surface points
Ul Haq, Wang, and Djurdjanovic [94]	Feature extraction from streaming signal data
Christ, Kempa-Liehr, and Feindt [95]	Feature extraction and selection from time series data
Wang et al. [99]	Clustering algorithms to extract representative data instances
Nikolaidis, Goulermas, and Wu [102]	Instance reduction based on distance from class boundaries

The second approach of interest is the use of optimization heuristics to obtain a near-optimal subset of features for the problem at hand. It might be a reasonable extension to TSFRESH to incorporate a second filter that seeks to better optimize the feature set with respect to the objective function, possibly using a heuristic such as simulated annealing. This would also add the dimension of feature interaction, which is currently not present in the TSFRESH statistical independence filter.

A final observation from the third category of reviewed literature is that the set of literature on reducing or filtering the features that might go into a machine learning model is reasonably robust but is relatively less robust concerning the prioritization of the remaining features. This implies a gap in terms of approaches to quantitatively or qualitatively stack features against each other. An alternative explanation is that such approaches exist but were simply not employed in the reviewed literature. This seems unlikely, as, the benefit of such capability would be to see how a particular feature of interest fares in its utility from one problem to the next. In smart manufacturing, the same features of data are continually collected and used repeatedly in different analyses. It may be of interest to know which of those features tend to be valuable in harboring predictive power and which ones tend not to.

4. Discussion

This paper reviewed existing research into frameworks or approaches for big-data analytics as applied to three levels of projects, with increasing degrees of precision or detail. The first level reviewed was a high-level look at frameworks for general industrial applications. The second level focused specifically and local, lower-level smart manufacturing applications of fault detection and fault prediction. Finally, the third and most specialized level looked at approaches specifically oriented towards data reduction.

In each section, articles were discussed individually as they pertain to the motivation for this research and their applicability to the CE. At the end of each section, discussion followed to summarize any observations across the set of articles within the section and relate them to each other. The final level of discussion is to look at the full picture and identify any observations, trends, or commonalities that span the three levels.

The first observation is that there is a dichotomy in how the same verbiage can be applied to different contexts. Terms like ‘framework’, ‘big data’, and ‘predictive analytics’ in some cases are contextualized as architecture required to build organizational competencies and in other cases as approaches or methodologies to build individual competencies.

In the context of organizational competencies for big data analytics, much consideration is made as to the architecture for where the data exists, how it moves from one location to the next, and at which level or echelon the analysis takes place. In general, there is some layer or module at which the initial data is generated or collected but then options for what to do with it. Data may be migrated to a cloud-based data center and analyzed in a consolidated location, or it may be analyzed at local nodes. Whether to analyze at the edge or in the cloud will typically be a function of resources and of the time window available to perform corrective action. Actions that require real-time processing for quick action might not be performed at the cloud level because, by the time data is captured, cleaned, pre-processed, and run through a model, the window to correct an identified fault may have already passed. On the other hand, if there is sufficient time between the data collection point and the decision point, such as a manufacturing process in which there might be a gap of hours or days between assembly line procedures and testing, then analysis at the cloud level might be suitable.

It should be noted that, from an organizational competency perspective, the infrastructure is a prerequisite to the development of individual competencies in the form of data scientist best practices. However, it is those data scientist best practices that become contributing factors to other organizational competencies such as knowledge management and decisions on long term data retention. There is an iterative and cyclic relationship such that organizational competencies produce individual competencies which then build and reinforce other organizational competencies.

A second observation across the three sections of reviewed literature is that there was a conspicuous absence of any discussion of the generalization of results beyond the specific problem of interest. This is true on both the 'front' end and the 'back' end of the articles reviewed. In other words, upon conclusion of the experiment or analysis, there was no discussion in any reviewed article of knowledge management or steps to generalize results from an input data perspective. There was certainly discussion about future research opportunities in generalizing an overall approach or algorithm, but in no cases did that discussion manifest itself the form of practical reflection on a feature set's utility for the problem of interest and prospects for utility in other scenarios. Similarly, during model formulation, there was no discussion of institutional knowledge that might play a role in feature selection. Only one reviewed paper referenced a data screening decision that was made based on prior work. The context in that situation was 21 possible signals to use as model inputs, of which seven were selected based on reviewed literature.

This observation is not intended as a negative criticism of any past work. It is quite natural to expect that this might be the case because finite resources drive priorities, and in a fast-paced world there is often little time to breathe between the completion of one project and the start of another. Given this reality, there appears to be value in anything that can facilitate the creation and preservation of institutional knowledge in this domain.

A third observation is that feature selection approaches in most cases were performed using a single technique at a single point in the model building process. Feature filtration using Kullback–Leibler divergence reduced features sets of 1460 and 1460 to 198 and 175, respectively. Feature filtration using Cohen's kappa was stated as a step in one case study, but no results were provided as to how many features were filtered out. A combination of PCA and subject matter expertise reduced a feature set of 1000 to 16, although it was not clearly identified how many of those features were reduced from PCA and how many from subject matter expertise. In the case of TSFRESH, statistical hypothesis tests for independence filter out features that are statistically independent, reducing 4764 features to 623.

A natural next step for any of these techniques is to explore the possibility to layer one technique after another depending on how many features remain after a given filter. In the case of 1000 features reduced to 16, it is possible to successively iterate through all 65536 subsets of features to arrive at an optimal subset with minimal effort. In the case of 4764 features reduced to 623, however, this is computationally impractical. It is unlikely that the optimal subset of the 623 remaining features would be all 623 of those features; a layered approach to continue to weed out features would be a value-added step to analyses with large numbers of features remaining. This is especially true if there is the desire for the model and its results to be understandable and digestible on the human side of the enterprise. Furthermore, what is understandable and digestible to the data scientist may be neither understandable nor digestible to the decision maker. Communication and visualization are critical components to the human element, particularly for decision makers who may not have background in the technical aspects of data science.

5. Conclusions

The papers reviewed are not intended to represent an exhaustive list of all existing research on the subject. However, it is believed that the reviewed examples do provide a representative sample of the sort of research currently performed in this discipline.

A conclusion that may be drawn from the first general observation in Section 4 is that there is value in having a standard set of terminology when speaking about the big data environment in order to distinguish when one is referring to organizational capabilities or individual competencies. In the reviewed literature, terms like 'framework' or 'data collection' carried wide variance in their meaning depending on the context. It is likely that standard terms will be settled on over time, either bottom-up from common use or top-down from professional organizations in industry, academia, or government. At this point, it may suffice simply to be aware of the different contexts in which the topic may be broached. Attention to detail is always a good rule of thumb in any endeavor, and that may be a good temporary solution for now.

A second conclusion, following from the second general observation in Section 4, is that a generalized approach to provide clarity as to what input data is valuable and what input data is not valuable, perhaps with both a quantitative and qualitative dimension, can shape analysis decisions in the big-data environment. Those decisions might be localized to the problem of interest, as in deciding which features to include in the model. Those decisions might also extend to larger, resource-oriented decisions, such as start-up priorities for transitioning from a legacy manufacturing facility to a CE. From a knowledge management standpoint, there is value in building institutional knowledge regarding features that perform poorly as well as features that perform well. Knowing which features tend to habitually appear in good solutions and which features habitually appear in bad solutions, if such knowledge exists, would be tremendously helpful in long term data capture and storage decisions.

Finally, the third general observation in Section 4 lends itself to the conclusion that there is room for additional research into practical means for feature filtration and prioritization. On the surface, there appears to be no reason why the single-layer filtration techniques employed in the reviewed articles cannot be extended into a series of hierarchical filters. One possible limitation would be that several of the techniques employ similarly-themed filters that may produce only limited improvement when performed in sequence. For example, filtering once by Cohen's kappa and then by testing for statistical independence might not produce substantial improvement. However, following the initial filtration by way of Cohen's kappa with an optimization heuristic such as simulated annealing or genetic algorithm to find an optimal or near-optimal subset of features might be a promising avenue to explore.

It is also worth exploring, from a knowledge management standpoint, feature reduction and selection techniques that preserve as much interpretability as possible. It has already been discussed that PCA is a common approach, but the reduction in dimensions from M to K, where K < M, will necessarily take away the physical meaning from those K features. Techniques in feature reduction that preserve the nature of the original features are a value-added contribution to this question.

In closing, manufacturing in the 21st century is a highly competitive enterprise, and the business value in exploiting 21st century technologies in smart manufacturing, IIOT, Industry 4.0, and the CE cannot be overstated. At the core of this opportunity for the individual manufacturer is the untapped potential held in the volumes of smart manufacturing data collected and stored in its data repositories. For an organization to develop as a core competency a methodological approach or process to build and continually develop institutional knowledge about the data landscape at its disposal, it would move the body of input data for any given predictive analytics project from simply a means to an end to an end unto itself. This is something of a paradigm shift, but one that can produce meaningful advantage to the organization that harnesses it.

Future research will develop and explore the potential for frameworks of this nature in the smart manufacturing context.

Author Contributions: Conceptualization, P.M.L., W.O, and F.P.M.; methodology, P.M.L.; software, N/A; validation, P.M.L., W.O. and F.P.M.; formal analysis, P.M.L.; investigation, P.M.L.; resources, F.P.M.; data curation, N/A; writing—original draft preparation, P.M.L.; writing—review and editing, P.M.L., W.O., F.P.M.; visualization, N/A.; supervision, W.O.; project administration, W.O.; funding acquisition, N/A.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kusiak, A. Smart manufacturing must embrace big data. *Nature* **2017**, *544*, 23–25. [CrossRef] [PubMed]
2. Tao, F.; Qi, Q.; Liu, A.; Kusiak, A. Data-driven smart manufacturing. *J. Manuf. Syst.* **2018**, *48*, 157–169. [CrossRef]
3. Everything You Need to Know about the Industrial Internet of Things. *G.E. Digital*. 2016. Available online: https://www.ge.com/digital/blog/everything-you-need-know-about-industrial-internet-things (accessed on 1 May 2018).
4. Schneider, S. The industrial internet of things (IIoT): Applications and taxonomy. In *Internet of Things and Data Analytics Handbook*; Wiley: Hoboken, NJ, USA, 2017; pp. 41–81.
5. Industrial Internet Consortium. Available online: https://www.iiconsortium.org/ (accessed on 2 May 2018).
6. OpenFog. Available online: https://www.openfogconsortium.org/ (accessed on 2 May 2018).
7. Lasi, H.; Fettke, P.; Kemper, H.G.; Feld, T.; Hoffmann, M. Industry 4.0. *Bus. Inf. Syst. Eng.* **2014**, *6*, 239–242. [CrossRef]
8. Gilchrist, A. *Industry 4.0: The Industrial Internet of Things*; Apress: New York, NY, USA, 2016.
9. Li, L. China's manufacturing locus in 2025: With a comparison of "Made-in-China 2025" and 'Industry 4.0'. *Technol. Forecast. Soc. Chang.* **2018**, *135*, 66–74. [CrossRef]
10. METI, Connected Industries. Ministry of Economy, Trade and Industry. 2017. Available online: http://www.meti.go.jp/english/policy/mono_info_service/connected_industries/index.html (accessed on 10 January 2019).
11. Granrath, L. Japan's Society 5.0: Going Beyond Industry 4.0. *Japan Industry News*. 2017. Available online: https://www.japanindustrynews.com/2017/08/japans-society-5-0-going-beyond-industry-4-0/ (accessed on 10 January 2019).
12. Rockwell Automation. *The Connected Enterprise eBook: Bringing People, Processes, and Technology Together*; Rockwell Automation: Milwaukee, WI, USA, 2015.
13. Otieno, W.; Cook, M.; Campbell-Kyureghyan, N. Novel approach to bridge the gaps of industrial and manufacturing engineering education: A case study of the connected enterprise concepts. In Proceedings of the 2017 IEEE Frontiers in Education Conference (FIE), Indianapolis, IN, USA, 18–21 October 2017; pp. 1–5.
14. Qin, S.J. Process data analytics in the era of big data. *AIChE J.* **2014**, *60*, 3092–3100. [CrossRef]
15. McKinsey & Company. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*; McKinsey Global Institute: Washington, DC, USA, 2011; p. 156.
16. Bollier, D.; Firestone, C.M. *The Promise and Peril of Big Data*; The Aspen Institute: Washington, DC, USA, 2010.
17. Lenz, J.; Wuest, T.; Westkämper, E. Holistic approach to machine tool data analytics. *J. Manuf. Syst.* **2018**, *48*, 180–191. [CrossRef]
18. Thoben, K.; Wiesner, S.; Wuest, T. 'Industrie 4.0' and Smart Manufacturing—A Review of Research Issues and Application Examples. *Int. J. Autom. Technol.* **2017**, *11*, 4–19. [CrossRef]
19. Kaufman, E.L.; Lord, M.W.; Reese, T.W.; Volkmann, J. The Discrimination of Visual Number. *Am. J. Psychol.* **1949**, *62*, 498–525. [CrossRef] [PubMed]

20. Miller, G.A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.* **1956**, *63*, 81–97. [CrossRef] [PubMed]
21. Simon, H.A. Designing organizations for an information-rich world. *Comput. Commun. Public Interes.* **1971**, *72*, 37.
22. Oussous, A.; Benjelloun, F.; Lahcen, A.A.; Belfkih, S. Big Data technologies: A survey. *J. King Saud Univ. Comput. Inf. Sci.* **2018**, *30*, 431–448. [CrossRef]
23. Honest, N. A Survey of Big Data Analytics. *Int. J. Inf. Sci. Tech.* **2016**, *6*, 35–43. [CrossRef]
24. Tsai, C.-W.; Lai, C.-F.; Chao, H.-C.; Vasilakos, A.V. Big data analytics: A survey. *J. Big Data* **2015**, *2*, 21. [CrossRef]
25. Spangenberg, N.; Roth, M.; Franczyk, B. Evaluating new approaches of big data analytics frameworks. In Proceedings of the International Conference on Business Information Systems, Poznań, Poland, 24–26 June 2015.
26. Wuest, T.; Weimer, D.; Irgens, C.; Thoben, K.-D. Machine learning in manufacturing: Advantages, challenges, and applications. *Prod. Manuf. Res.* **2016**, *4*, 23–45. [CrossRef]
27. Dingli, D.J. *The Manufacturing Industry—Coping with Challenges*; Working Paper No. 2012/05; 2012; p. 47. Available online: https://econpapers.repec.org/paper/msmwpaper/2012_2f05.htm (accessed on 27 February 2018).
28. Gordon, J.; Sohal, A.S. Assessing manufacturing plant competitiveness—An empirical field study. *Int. J. Oper. Prod. Manag.* **2001**, *21*, 233–253. [CrossRef]
29. Shiang, L.E.; Nagaraj, S. Impediments to innovation: Evidence from Malaysian manufacturing firms. *Asia Pac. Bus. Rev.* **2011**, *17*, 209–223. [CrossRef]
30. Thomas, A.J.; Byard, P.; Evans, R. Identifying the UK's manufacturing challenges as a benchmark for future growth. *J. Manuf. Technol. Manag.* **2012**, *23*, 142–156. [CrossRef]
31. Kotsiantis, S.B. Supervised Machine Learning: A Review of Classification Techniques. *Informatica* **2007**, *31*, 249–268.
32. Yang, K.; Trewn, J. *Multivariate Statistical Methods in Quality Management*; McGraw-Hill: New York, NY, USA, 2004.
33. Alpaydin, E. *Introduction to Machine Learning*, 3rd ed.; MIT Press: Cambridge, MA, USA, 2014.
34. Doltsinis, S.; Ferreira, P.; Lohse, N. Reinforcement learning for production ramp-up: A Q-batch learning approach. In Proceedings of the 11th International Conference on Machine Learning and Applications, Boca Raton, FL, USA, 12–15 December 2012; pp. 610–615.
35. Wang, J.; Ma, Y.; Zhang, L.; Gao, R.X.; Wu, D. Deep learning for smart manufacturing: Methods and applications. *J. Manuf. Syst.* **2018**, *48*, 144–156. [CrossRef]
36. Butler, B. What Is Edge Computing and How It's Changing the Network. *Network World*. 2017. Available online: https://www.networkworld.com/article/3224893/internet-of-things/what-is-edge-computing-and-how-it-s-changing-the-network.html (accessed on 7 March 2018).
37. Linthicum, D. Responsive Data Architecture for the Internet of Things. *Computer* **2016**, *49*, 72–75. [CrossRef]
38. Mahmud, R.; Kotagiri, R.; Buyya, R. Fog Computing: A Taxonomy, Survey and Future Directions. In *Internet of Everything;* Springer: Singapore, 2018; pp. 103–130.
39. Flath, C.M.; Stein, N. Towards a data science toolbox for industrial analytics applications. *Comput. Ind.* **2018**, *94*, 16–25. [CrossRef]
40. Kumar, A.; Shankar, R.; Choudhary, A.; Thakur, L.S. A big data MapReduce framework for fault diagnosis in cloud-based manufacturing. *Int. J. Prod. Res.* **2016**, *54*, 7060–7073. [CrossRef]
41. Japkowicz, N.; Stephen, S. The class imbalance problem: A systematic study. *Intell. Data Anal.* **2002**, *6*, 429–449. [CrossRef]
42. Longadge, R.; Dongre, S.S.; Malik, L. Class imbalance problem in data mining: Review. *Int. J. Comput. Sci. Netw.* **2013**, *2*, 83–87.
43. Bahga, A.; Madisetti, V.K. Analyzing massive machine maintenance data in a computing cloud. *IEEE Trans. Parallel Distrib. Syst.* **2012**, *23*, 1831–1843. [CrossRef]
44. Devaney, M.; Cheetham, B. Case-Based Reasoning for Gas Turbine Diagnostics. In Proceedings of the 18th International FLAIRS Conference (FLAIRS-05), Clearwater Beach, FL, USA, 16–18 May 2005.
45. Timmerman, H. SKF WindCon Condition Monitoring System for Wind Turbines. In Proceedings of the New Zealand Wind Energy Conference, Wellington, NZ, USA, 20–22 April 2009.

46. Tamilselvan, P.; Wang, P. Failure diagnosis using deep belief learning based health state classification. *Reliab. Eng. Syst. Saf.* **2013**, *115*, 124–135. [CrossRef]
47. Hinton, G.E. A Practical Guide to Training Restricted Boltzmann Machines. *Computer* **2012**, *9*, 599–619.
48. Jia, F.; Lei, Y.; Lin, J.; Zhou, X.; Lu, N. Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. *Mech. Syst. Signal Process.* **2016**, *72–73*, 303–315. [CrossRef]
49. Schmidhuber, J. Deep Learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef] [PubMed]
50. Banerjee, T.; Das, S.; Roychoudhury, J.; Abraham, A. Implementation of a New Hybrid Methodology for Fault Signal Classification Using Short-Time Fourier Transform and Support Vector Machines. In Proceedings of the 5th International Workshop on Soft Computing Models in Industrial Environment Application (SOCO 2010), Guimarães, Portugal, 16–18 June 2010; Volume 73, pp. 219–225.
51. Banerjee, T.P.; Das, S. Multi-sensor data fusion using support vector machine for motor fault detection. *Inf. Sci.* **2012**, *217*, 96–107. [CrossRef]
52. Jack, L.B.; Nandi, A.K. Fault detection using support vector machines and artificial neural networks, augmented by genetic algorithms. *Mech. Syst. Signal Process.* **2002**, *16*, 373–390. [CrossRef]
53. Rychetsky, M.; Ortmann, S.; Glesner, M. Support vector approaches for engine knock detection. In Proceedings of the IJCNN'99. International Joint Conference on Neural Networks, Washington, DC, USA, 10–16 July 1999; Volume 2.
54. Altintas, Y. In-process detection of tool breakages using time series monitoring of cutting forces. *Int. J. Mach. Tools Manuf.* **1988**, *28*, 157–172. [CrossRef]
55. Wang, H.; Zhoui, J.; He, I.; Sha, J. An uncertain information fusion method for fault diagnosis of complex system. In Proceedings of the 2003 International Conference on Machine Learning and Cybernetics, Xi'an, China, 5 November 2003; pp. 1505–1510.
56. Xiong, J.; Zhang, Q.; Sun, G.; Zhu, X.; Liu, M.; Li, Z. An Information Fusion Fault Diagnosis Method Based on Dimensionless Indicators with Static Discounting Factor and KNN. *IEEE Sens. J.* **2016**, *16*, 2060–2069. [CrossRef]
57. Dempster, A.P. A Generalization of Bayesian Inference. *J. R. Stat. Soc.* **1968**, *30*, 205–247. [CrossRef]
58. Khakifirooz, M.; Chien, C.F.; Chen, Y.J. Bayesian inference for mining semiconductor manufacturing big data for yield enhancement and smart production to empower industry 4.0. *Appl. Soft Comput. J.* **2017**, *68*, 990–999. [CrossRef]
59. Lee, H. Framework and development of fault detection classification using IoT device and cloud environment. *J. Manuf. Syst.* **2017**, *43*, 257–270. [CrossRef]
60. Gunes, V.; Peter, S.; Givargis, T.; Vahid, F. A Survey on Concepts, Applications, and Challenges in Cyber-Physical Systems. *KSII Trans. Internet Inf. Syst.* **2014**, *8*, 120–132.
61. Rajkumar, R.; Lee, I.; Sha, L.; Stankovic, J. Cyber-physical systems. In Proceedings of the 47th Design Automation Conference on—DAC '10, Anaheim, CA, 13–18 June 2010; p. 731.
62. Saez, M.; Maturana, F.; Barton, K.; Tilbury, D. Modeling and Analysis of Cyber-Physical Manufacturing Systems for Anomaly Detection and Diagnosis. 2018. Available online: https://www.nist.gov/sites/default/files/documents/2018/05/22/univ_michigan_miguel_saez.pdf (accessed on 25 February 2019).
63. Saez, M.; Maturana, F.; Barton, K.; Tilbury, D. Anomaly detection and productivity analysis for cyber-physical systems in manufacturing. In Proceedings of the 2017 13th IEEE Conference on Automation Science and Engineering (CASE), Xi'an, China, 20–23 August 2017; pp. 23–29.
64. Wan, J.; Tang, S.; Li, D.; Wang, S.; Liu, C.; Abbas, H.; Vasilakos, A.V. A Manufacturing Big Data Solution for Active Preventive Maintenance. *IEEE Trans. Ind. Inform.* **2017**, *13*, 2039–2047. [CrossRef]
65. Munirathinam, S.; Ramadoss, B. Big data predictive analtyics for proactive semiconductor equipment maintenance. In Proceedings of the 2014 IEEE International Conference on Big Data (IEEE Big Data 2014), Washington, DC, USA, 27–30 October 2014; pp. 893–902.
66. Franklin, J. Signalling and anti-proliferative effects mediated by gonadotrophin-releasing hormone receptors after expression in prostate cancer cells using recombinant adenovirus. *J. Endocrinol.* **2003**, *176*, 275–284. [CrossRef] [PubMed]
67. Ji, W.; Wang, L. Big data analytics based fault prediction for shop floor scheduling. *J. Manuf. Syst.* **2017**, *43*, 187–194. [CrossRef]

68. Rolfe, B.F.; Frayman, Y.; Kelly, G.L.; Nahavandi, S. Recognition of Lubrication Defects in Cold Forging Process with a Neural Network. In *Artificial Neural Networks in Finance and Manufacturing*; IGI Global: Hershey, PA, USA, 2006; pp. 262–275.
69. Perzyk, M.; Kochański, A.W. Prediction of ductile cast iron quality by artificial neural networks. *J. Mater. Process. Technol.* **2001**, *109*, 305–307. [CrossRef]
70. Kilickap, E.; Yardimeden, A.; Çelik, Y.H. Mathematical Modelling and Optimization of Cutting Force, Tool Wear and Surface Roughness by Using Artificial Neural Network and Response Surface Methodology in Milling of Ti-6242S. *Appl. Sci.* **2017**, *7*, 1064. [CrossRef]
71. Huang, C.; Jia, X.; Zhang, Z. A modified back propagation artificial neural network model based on genetic algorithm to predict the flow behavior of 5754 aluminum alloy. *Materials* **2018**, *11*, 855. [CrossRef] [PubMed]
72. Arnaiz-González, Á.; Fernández-Valdivielso, A.; Bustillo, A.; de Lacalle, L.N.L. Using artificial neural networks for the prediction of dimensional error on inclined surfaces manufactured by ball-end milling. *Int. J. Adv. Manuf. Technol.* **2016**, *83*, 847–859. [CrossRef]
73. De Lacalle, L.N.L.; Lamikiz, A.; Salgado, M.A.; Herranz, S.; Rivero, A. Process planning for reliable high-speed machining of moulds. *Int. J. Prod. Res.* **2002**, *40*, 2789–2809. [CrossRef]
74. De Lacalle, L.N.L.; Lamikiz, A.; Sánchez, J.A.; Salgado, M.A. Effects of tool deflection in the high-speed milling of inclined surfaces. *Int. J. Adv. Manuf. Technol.* **2004**, *24*, 621–631. [CrossRef]
75. Lasemi, A.; Xue, D.; Gu, P. Recent development in CNC machining of freeform surfaces: A state-of-the-art review. *CAD Comput. Aided Des.* **2010**, *42*, 641–654. [CrossRef]
76. Liu, X.; Li, Y. Feature-based adaptive machining for complex freeform surfaces under cloud environment. *Robot. Comput. Integr. Manuf.* **2019**, *56*, 254–263. [CrossRef]
77. García, S.; Luengo, J.; Herrera, F. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowl.-Based Syst.* **2015**, *98*, 1–29. [CrossRef]
78. Liu, H.; Setiono, R. A Probabilistic Approach to Feature Selection—A Filter Solution. In Proceedings of the Thirteenth International Conference on Machine and Learning, Bari, Italy, 3–6 July 1996; pp. 319–327.
79. Battiti, R. Using Mutual Information for Selecting Features in Supervised Neural-Net Learning. *IEEE Trans. Neural Netw.* **1994**, *5*, 537–550. [CrossRef] [PubMed]
80. Kira, K.; Rendell, L. A practical approach to feature selection. In Proceedings of the Ninth International Conference on Machine Learning, Aberdeen, UK, 1–3 July 1992; pp. 249–256.
81. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef] [PubMed]
82. Hart, P. The condensed nearest neighbor rule (Corresp.). *IEEE Trans. Inf. Theory* **1968**, *14*, 515–516. [CrossRef]
83. Wilson, D.L. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Trans. Syst. Man Cybern.* **1972**, *2*, 408–421. [CrossRef]
84. Wilson, D.R.; Martinez, T.R. Reduction Techniques for Instance-Based Learning Algorithms. *Mach. Learn.* **2000**, *38*, 257–286. [CrossRef]
85. Brighton, H.; Mellish, C. Advances in Instance Selection for Instance-Based Learning Algorithms. *Data Min. Knowl. Discov.* **2002**, *6*, 153–172. [CrossRef]
86. Stanula, P.; Ziegenbein, A.; Metternich, J. Machine learning algorithms in production: A guideline for efficient data source selection. *Procedia CIRP* **2018**, *78*, 261–266. [CrossRef]
87. Rehman, M.H.U.; Chang, V.; Batool, A.; Wah, T.Y. Big data reduction framework for value creation in sustainable enterprises. *Int. J. Inf. Manag.* **2016**, *36*, 917–928. [CrossRef]
88. Luan, T.H.; Gao, L.; Li, Z.; Xiang, Y.; Wei, G.; Sun, L. Fog Computing: Focusing on Mobile Users at the Edge. *arXiv*, 2015; arXiv:1502.01815.
89. Ma, X.; Cripps, R.J. Shape preserving data reduction for 3D surface points. *CAD Comput. Aided Des.* **2011**, *43*, 902–909. [CrossRef]
90. Jeong, I.-S.; Kim, H.-K.; Kim, T.-H.; Lee, D.H.; Kim, K.J.; Kang, S.-H. A Feature Selection Approach Based on Simulated Annealing for Detecting Various Denial of Service Attacks. *Converg. Secur.* **2016**, *2016*, 1–18. [CrossRef]
91. Kang, S.-H.; Kim, K.J. A feature selection approach to find optimal feature subsets for the network intrusion detection system. *Cluster Comput.* **2016**, *19*, 325–333. [CrossRef]

92. Du, K.L.; Swamy, M.N.S. *Search and Optimization by Metaheuristics: Techniques and Algorithms Inspired by Nature*; Springer: Basel, Switzerland, 2016; pp. 1–434.
93. Lalehpour, A.; Berry, C.; Barari, A. Adaptive data reduction with neighbourhood search approach in coordinate measurement of planar surfaces. *J. Manuf. Syst.* **2017**, *45*, 28–47. [CrossRef]
94. Haq, A.A.U.; Wang, K.; Djurdjanovic, D. Feature Construction for Dense Inline Data in Semiconductor Manufacturing Processes. *IFAC-PapersOnLine* **2016**, *49*, 274–279. [CrossRef]
95. Christ, M.; Kempa-Liehr, A.W.; Feindt, M. Distributed and parallel time series feature extraction for industrial big data applications. *arXiv*, 2016; arXiv:1610.07717.
96. Benjamini, Y.; Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **2001**, *29*, 1165–1188.
97. Dheeru, D.; Taniskidou, E.K. *UCI Machine Learning Repository*; School of Information and Computer Sciences, University of California: Irvine, CA, USA, 2017.
98. Christ, M.; Braun, N.; Neuffer, J.; Kempa-Liehr, A.W. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh—A Python package). *Neurocomputing* **2018**, *307*, 72–77. [CrossRef]
99. Wang, W.; Liu, J.; Pitsilis, G.; Zhang, X. Abstracting massive data for lightweight intrusion detection in computer networks. *Inf. Sci.* **2018**, *433–434*, 1339–1351. [CrossRef]
100. Fan, J.; Han, F.; Liu, H. Challenges of Big Data analysis. *Natl. Sci. Rev.* **2014**, *1*, 293–314. [CrossRef] [PubMed]
101. Campos, J.; Sharma, P.; Gabiria, U.G.; Jantunen, E.; Baglee, D. A Big Data Analytical Architecture for the Asset Management. *Procedia CIRP* **2017**, *64*, 369–374. [CrossRef]
102. Nikolaidis, K.; Goulermas, J.Y.; Wu, Q.H. A class boundary preserving algorithm for data condensation. *Pattern Recognit.* **2011**, *44*, 704–715. [CrossRef]