Abstract: Virtual sensors, or soft sensors, have greatly contributed to the evolution of the sensing systems in industry. The soft sensors are process models having three fundamental categories, namely white-box (WB), black-box (BB) and gray-box (GB) models. WB models are based on process knowledge while the BB models are developed using data collected from the process. The GB models integrate the WB and BB models for addressing the concerns, i.e., accuracy and intuitiveness, of industrial operators. In this work, various design aspects of the GB models are discussed followed by their application in the process industry. In addition, the changes in the data-driven part of the GB models in the context of enormous amount of process data collected in Industry 4.0 are elaborated.

Keywords: big data analytics; internet of things; machine learning; sensor 4.0

1. Introduction

Industrial evolution in terms of automation and control has been classified into four major eras, i.e., Industry 1.0, 2.0, 3.0 and 4.0. The introduction of mechanical production facilities led to the first industrial revolution termed as Industry 1.0 which span through an era from the second half of the eighteenth century till the last quarter of the nineteenth century. The second industrial revolution, termed Industry 2.0, started in the 1870s on the emergence of electricity applications and proper management of labor. The digitization introduced around the 1970s helped use advanced electronics and information technology for process automation, which formed the era of Industry 3.0. The current (r)evolution, termed Industry 4.0, is driven by the emergence of concepts, such as cloud-computing, smart sensors, internet of things (IoT), big data analytics, augmented reality, and human–machine interfaces [1–8].

Process sensors also evolved from purely mechanical indicators used in the era of Industry 1.0 to smart sensors of Industry 4.0. In this context, Sensors 4.0 has been coined colloquially to Industry 4.0 [9]. Smart sensors are sensing devices that are equipped with digital features for data processing, storage and efficient transformation of data [10–12]. Smart sensors automatized the zero correction, calibration, and scaling of the measured signals by using microprocessors, in contrast to meticulous design, testing, and debugging faced by the conventional analog sensors [13–16]. The sensing platform of a smartphone is the best example of smart sensing environment; it integrates multiple sensors and
makes use of multi-sensory signals based on gyroscopes, accelerators, pressure, and magnetometers sensors. These sensors simultaneously do weather monitoring, step counting, screen orientation, gaming, etc.

The use of smart sensors in the process industry has been demonstrated in several studies [17–21], where internet of things (IoT) devices are used to collect enormous amount of data from process equipment such hydraulic pumps, reactors, heat exchangers, etc. Considering the massive amount of data, data processing is performed before it is passed on for analysis and decision making regarding the process operations.

The data processing methods have to handle the big volume, high frequency, variety and veracity of the data coming from the IoT devices [22,23]. The preprocessing techniques include data wrangling, visualization, sparsity and regularization, optimization, reducing dimensionality, measuring distance, representation learning, and sequential learning [24–27].

The data collected through the hardware sensors, i.e., thermo-couples, manometers, etc., have been used for monitoring and control of the process. However, three decades ago, the idea of using the data for the development of predictive models emerged [28]. These predictive models were termed soft sensors. The term soft sensor is a combination of two words; “software”, which refers to the fact that the models are usually computer programs, and “sensors” because they serve the same purpose as that of the hardware sensors. Soft sensors are capable of estimating process states that are difficult to measure through hardware sensors due to large measurement delays, technological limitations or high investment costs. Even if hardware sensors can be used, operators have found soft sensors superior to hardware sensors; the soft sensors stabilize operation, reduce energy and materials consumption, and cross-check the performance of an online hardware sensor [29]. Soft sensors can also be used in parallel with the hardware sensor to give additional information [30,31]. The soft sensors are classified into three major categories based on their underlying models namely white-box (WB), black-box (BB) and gray-box (GB) models. WB models are based on process knowledge while the BB models are developed using data collected from the process. The WB models are descriptive and easy for industry operators to interpret. However, they have certain limitations, sometimes they are unable to grasp the true dynamics of a complex industrial process and their prediction accuracy is accompanied by a substantial amount of errors. The BB models, on the other hand, are more accurate in prediction than the WB models, however, they are less intuitive in nature. The GB models integrate the WB and BB models for addressing the concerns, i.e., accuracy and intuitiveness, of industrial operators.

Applications of GB models are found in a variety of domains in process industries such as design, estimation, control, and monitoring. They are applied across the process industries such as iron and steel making [32–36], food processing [37–39], oil and gas processing [40–44], chemical, biochemical, and pharmaceutical [45–54], power plants [55–57], water treatment [58–60], material processing and energy materials [61–64] and industrial robots [65–67].

The aim of the current study is to do a comprehensive review of the fundamentals of GB soft sensors, their applications in the process industry and prospects in the era of big data analytics; to the best of the authors’ knowledge, no extensive review that addresses all such dimensions is reported in the literature. GB soft sensors application in selected process industries were investigated. The industries were ranked in terms of using GB soft sensors in their process. The pattern in the objectives of the application, the trends in the type of GB model as well as the type of BB methods were analyzed. Finally, the prospects and challenges of GB soft sensors in the era of big data are elaborated.

Section 2 describes fundamentals of soft sensors. The design types of GB models and methodology adopted in this paper are discussed in Sections 3 and 4, respectively. Applications of GB models in the process industry and their prospects in the era of industry 4.0 are discussed in Section 5 followed by conclusions in Section 6.
2. Fundamentals of Soft Sensors

The fundamentals of three types of soft sensors based on their underlying model, i.e., WB, BB, and GB, are briefly discussed in this section.

The WB models are also referred to as the first principle, “mechanistic”, “analytical”, “phenomenological”, “physical”, “fundamental”, and “parametric” models that describe the underlying laws of science and engineering which govern the process(es) [68]. The WB models rely on natural characteristics, i.e., reaction kinetics, thermodynamics, and fluid properties, and conservation laws, such as mass, energy, and momentum of the target process [69]. The WB models transform the process of knowledge into mathematical formulations [70]. The WB model equations get a form of ordinary or partial differential-algebraic equations with properly defined initial and boundary conditions. Analytical methods or numerical methods are applied to solve the equations keeping in view the complexity of the problem. Some parameters of the WB models can be based on real data to a minor extent [71,72]. For example, Fick’s law, Fourier’s law, Darcy’s law, and ideal gas law are categorized into WB models, but they were initially derived through empirical correlations, based on experimental data. The attempts to efficiently develop process models through simulator, a WB modeling environment, started back in the 1950s. However, the emergence of Process Systems Engineering later in the 1980s intensified the integration of process simulations in the loop of computer-aided design, process control, and optimization [73,74]. The WB models face some challenges such as non-linearity [75,76], uncertainties [77,78], multi-scale (time and physical dimensions) [79], high dimensionality [80], and time delay [81].

The BB models are used to describe complex processes that are difficult for the WB model to handle [82]. Other names used for BB models are “data-based”, “non-parametric”, and “empirical” models [68]. The BB term refers to black-box behaviors of merely mapping the I/O data and its mathematical structure is not necessarily based on natural characteristics of the process [83]. The lack of physical meaning of the BB model structures is considered its disadvantage but this feature also makes the researchers able to model a process in spite of their unawareness of underlying process dynamics [83,84]. With increasing complexities of modeling tasks, emergence of high-speed computing and demand of less complicated and efficient process sensing, a paradigm shift from WB to BB methods, such as artificial neural network (ANN), partial least squares (PLS), support vector machine (SVM), principle component analysis (PCA), and random forests (RF) have been applied [28,82,85]. However, the BB models are expensive in terms of the computational load and time than the WB models. Besides, realizing optimal structure design, accurate parametric values, and lesser intuitiveness have been the major challenges they face [86].

The GB models which emerged in the field of control and system theory in the 1990s combine the WB and the BB techniques [87–89]. The GB models are also named “semi-analytical”, “semi-physical”, “semi-parametric”, and “hybrid” models [68]. The hybrid term also represents the integration of two BB models. Thus, all GB models can be referred to as hybrid models, but all hybrid models may not be GB models. The GB models have been used in modeling a variety of processes since its inception [90,91]. GB modeling strategy compensates for the deficiencies of the standalone WB and BB models by adding both accuracy, reliability, and intuitiveness [90].

3. Types of GB Models

GB models are classified into three categories: parallel, serial and combined gray-box models. A schematic view of GB models’ classification is shown in Figure 1.

The parallel GB model uses BB model to compensate error of a WB model. The BB model in the parallel GB modeling framework is referred as the outer statistical model as it does not affect the internal structure of a WB model. Likewise, the parallel GB approach overcomes the limitations caused
by the structure of the first-principle model. The general format of parallel GB mode is represented by the following equation [34]:

\[ \hat{y}_{pa} = f_{fp}(x_{fp}, \hat{\theta}) + f_{pa}(x, \phi_{pa}), \]  

(1)

where \( \hat{\theta} \) and \( \phi_{pa} \) are vectors of parameters of the WB and BB models, respectively. The \( x_{fp} \) is a subset of \( x \), and \( \hat{y}_{pa} \) is the prediction of \( y \) by using the parallel GB model.

In the parallel GB model, the error of the WB model is compensated, however, parameters of the WB model are kept constant. This simplification sometimes deteriorates the performance of the WB model because some parameters may strongly depend on process conditions. Therefore, the serial GB model is used to update the parameters as functions of process conditions. The serial GB models have higher intuitiveness because the important physical parameters are identified and related to process variables. The serial GB model is represented by the following equation:

\[ \hat{y}_{se} = f_{fp}(x_{fp}, \hat{\theta}_{ci}, \hat{\phi}_{com}), \]  

(2)

where \( \hat{\theta}_{ci} \) is the vector of the estimated parameters, and \( \hat{y}_{se} \) is the prediction of \( y \) by the serial GB model.

In order to utilize both the internal and external use of the BB models with the WB models, a combined GB model is developed by combing the parallel GB modeling and the serial GB modeling features. In the combined GB model, a prediction error of a serial GB model is compensated by an outer BB model. The combined GB model consists of the WB model, the inner BB model to estimate parameters, and the outer BB model to compensate the prediction error. The combined GB model is represented by the following equation.

\[ \hat{y}_{com} = f_{fp}(x_{fp}, \hat{\theta}_{ci}, \hat{\phi}), + f_{com}(x, \phi_{com}), \]  

(3)

where \( \hat{\phi}_{com} \) is a vector of parameters and \( \hat{y}_{com} \) is the prediction of \( y \) through the combined GB model.

![Figure 1. Generalized framework grey-box (GB) modeling [34].](image)

4. Methodology

Application of GB soft sensors in iron and steel making, food processing, oil and gas processing, chemical, biochemical, and pharmaceutical, power plants, and sub-processes such as water treatment, material processing, energy materials, and industrial robots were analyzed. In addition, application of GB soft sensors to miscellaneous theoretical case studies of process equipment were investigated.
These case studies were lumped into two categories, i.e., reactive systems and heating systems. Google scholar database was used to collect reported studies on GB soft sensors application in the selected process industries in the last fifteen years, i.e., 2005-2019. The industries were ranked in terms of the use of GB soft sensors in their process. The pattern in the objectives, i.e., quality estimation, fault diagnosis, control, etc., of using the GB soft sensors was also investigated. In addition, the type of GB as well as the type of BB methods, i.e., ANN, RF, SVM, etc., were also ranked in terms of their application. Finally, the prospects and challenges of GB soft sensors in the era of big data are elaborated.

5. Current Practice in Process Industry

5.1. Iron and Steelmaking

Sohlberg et al. [32] developed a serial GB model for estimation of exit concentration of hydrochloric acid of the pickling process. Taylor series expansion was used in integration with WB model of the process. In a study by Barrios et al. [33], a parallel GB model was developed for entry temperature estimation of secondary scale breaker (SB). Fuzzy Inference System (FIS) was used for the estimation of the error of the WB model of the process. Okura et al. [92] used a parallel GB model for the estimation of molten steel temperature in a continuous casting process. Partial least squares (PLS) and random forests (RF) were used as BB models. Ahmad et al. [34] developed parallel, serial, and combined GB models to predict and control molten steel temperature in a continuous casting process where RF was used as a BB model. In another study by Ahmad et al. [35], a combined GB model was integrated with a bootstrap filter to predict the probability distribution of molten steel temperature in a continuous casting process under uncertainty. Barrios et al. [36] used ANN-based parallel GB model to estimate scale breaker entry temperature.

5.2. Food Processing

Cubillos et al. [37] developed a serial GB for estimation and control of moisture content in a direct fish-meal rotary dryer where ANN was used as a BB model. Vieira et al. [38] developed a serial GB model for prediction and control of the moisture content of milk powder produced in a spouted bed dryer. ANN was used as BB technique. Saltık et al. [39] used a serial GB for estimation of membrane fouling for an ultrafiltration membrane unit in a whey separation process. Exponential static membrane resistance function was used for the identification of parameters of the WB model of the process.

5.3. Chemical, Biochemical, and Pharmaceutical

Prada-Moraga et al. [45] developed a serial GB for estimation of the growth rate of biomass in a fermentation process. A mixed-integer optimization algorithm Automatic Learning of Algebraic Models (ALAMO) was used to identify the structure and parameters of the model. Wang et al. [46] applied serial GB model of marine alkaline protease fermentation to predict biomass concentration, substrate concentration and relative enzyme activity. Multi-I/O least squares support vector machine (MLSSVM) integrated with an artificial bee colony optimization algorithm was used as a black-box model. Niu et al. [47] applied a parallel GB model for the prediction of substrate concentration, cell concentration, and product concentration of nosiheptide fed-batch fermentation. Least-squares support vector machines were used as a BB model for compensation of error of the WB model of the process. Wu et al. [93] developed a parallel GB method for modeling and control of polymer molecular weight distribution (MWD) of the petrochemical industry. A recurrent neural network (RNN) and an orthogonal polynomial feed-forward neural network (OPFNN) were combined to model the shape of MWD. Johansen et al. [48] investigated the use of a serial GB model to design a control scheme for screw speed of plasticating twin-screw extruder (TSE) of a petrochemical plant. Autoregressive moving average exogenous (ARMAX) structure was used as a BB part of the GB model. Liu et al. [49] proposed a serial GB modeling for prediction and control of melt viscosity of product of a polymer extrusion process of a petrochemical plant. Genetic algorithm was integrated.
with the WB model to develop the GB model. Everett et al. [50] developed a serial GB model to predict the non-linear behavior of mold cooling. ANN was used to approximate the parameters of state-space model (SSM) of the process. Zahedi et al. [51] used a serial GB model for yield prediction of an extraction process. A neuro-fuzzy technique was used as a black box model for the estimation of Sh number. Cubillos et al. [52] investigated the feasibility of using a serial GB model for Real Time Optimization (RTO) of a Williams-Otto reactor. ANN and GA were integrated with the first-principle model of the reactor to efficiently predict and control the reactor temperature and the flow rate of the component. Pitarc et al. [53] proposed a serial GB model for prediction of overall heat transfer coefficient of heat exchangers in an evaporation plant. Data reconciliation (DR) and polynomial constrained regression approaches were used as BB models. Liu et al. [54] used a serial GB model for estimation of mycelia concentration, sugar concentration and chemical potency of the fermentation process of the pharmaceutical industry where ANN was used as a BB model.

5.4. Power Plants

Zhao et al. [55] developed a serial GB model for estimation of boiler thermal efficiency and NOx concentration of furnace outlet in a coal power plant. Boiler thermal efficiency and NOx concentration of furnace outlet were the outputs of the model. The fast recursive algorithm was used as a BB model. Arahal et al. [56] used a serial GB model to predict and control the temperature of a thermal storage tank of a solar power plant. Simultaneous Perturbation Stochastic Approximation (SPSA) technique was used for parameter identification. Barszcż et al. [57] used a serial GB approach for anomaly detection and control of feed water conditions, i.e., temperature, pressure, flow rate, etc., of a heat exchanger of a coal power plant. ANN was used as a BB model.

5.5. Oil and Gas Processing

Møller et al. [40] investigated the use of a serial GB model for slugging oscillation of a valve control system in the offshore process. Extended Kalman Filter (EKF) was used as a BB model. Onel et al. [41] used a serial GB model for prediction of reactor output composition of steam methane reforming (SMR) microchannel reactor. The non-linear fitting model was used as a BB model. Bram et al. [42] developed a serial GB model to estimate and control the rejected flow rate of a hydro cyclone. The model parameters were estimated through a least-squares method. In a study by Lotfalipour et al. [43], a serial GB model was developed for prediction of CO2 emission in plant-wide facility of oil and gas rig. The least-squares estimation sequence is used as BB technique. Durrani et al. [44] used a serial GB strategy that was devised by integrating ANN with Aspen HYSYS model of crude distillation unit (CDU). ANN was used to estimate parameters, i.e., cut-point temperature, of the CDU model.

5.6. Water Treatment

Stentoft et al. [58] developed a serial GB model for prioritizing aeration in economical schedule in wastewater treatment plants. The genetic algorithm (GA) was integrated with stochastic differential equations, based on the WB model of the process. Stentoft et al. [59] used the serial modeling framework to predict ammonium nitrate concentrations in a small recirculating Water Resource Recovery Facilities (WRRFs). Extended Kalman Filter (EKF) was used for parameters estimation. Dragoi et al. [60] used serial, parallel and combined GB models to predict dissolution rate of solute, i.e., sodium carbonate, urea, and sodium bicarbonate. ANN was used as BB model.

5.7. Material Processing and Energy Materials

Li et al. [61] developed a serial GB model for estimating the possibility of the datum in materials authorization (RMA) process in a TFT-LCD industry. The fuzzy membership function (MF) value was used as a representative of the possibility of the datum. Ordinary least square method was used for parameters estimation. Rad et al. [62] applied a serial GB model for prediction and control of temperature of a thermostatic system. Simulink Design Optimization Tool (SDO) was used
for parameter identification. Masoudinejad et al. [63] investigated the use of a serial GB model of a photovoltaic (PV) cell for low illuminance indoor lighting conditions in a material handling and warehousing. Internal parameters were identified using a least squares method. Liu et al. [64] used a serial GB model for estimating the irradiation angle of sensor prototype in solar energy harvesting industrial facility. A simple least-squares method was used to predict the parameters of the WB model.

5.8. Industrial Robot

Wernholt et al. [65] developed a serial GB model for prediction of motor angular speed of an industrial robot. Weighted logarithmic least squares were used as a BB model. Knoblach et al. [94] used a serial GB model for prediction and control of the motor velocity of an industrial robot. Weighted logarithmic least squares (WLLS) was used as a BB model. Ayala et al. [66] used a serial GB for prediction of the deflection of a piezoelectric micromanipulator through ANN with data acquired in a laboratory setup. Wernholt et al. [67] investigated the use of GB model for prediction and control of the machine position of a robot. Weighted nonlinear least squares and weighted logarithmic least squares were used for parameter estimation.

5.9. Miscellaneous

5.9.1. Reactive Systems

In a study by Acuña et al. [95], Least-Square Support Vector Machine (LS-SVM) was used to develop a serial GB model for prediction of the degree of progress of reaction of a Continuous Stirred Tank Reactor (CSTR). In another study by Acuña et al. [96], serial GB model was developed for prediction of the degree of progress of the reaction in CSTR. Least-square support vector machine and genetic algorithms were used for enhancing the performance of the WB model of the CSTR. Porru et al. [97] developed a serial GB model for prediction of product composition of a heterogeneous gas-solid reactor. ANN and extended Kalman filter (EKF) was integrated with the WB model of the reactor. Xiong et al. [98] investigated the use of a parallel GB model for prediction and control of heat release inside a simulated exothermic batch reactor. ANN was used for compensation of error of the WB model of the reactor. Hourfar et al. [99] used a serial GB model that was developed for prediction and control of the temperature of the nonlinear CSTR benchmark process. ANN was used to estimate parameters of the WB model of the CSTR. Zanardo et al. [100] developed a serial GB model for prediction of molar flow rates of NOx and NH3 at the outlet of a Selective Catalytic Reduction (SCR). Auto-regressive with eXogenous Input (ARXIs) was used for parameter identification. Acuña et al. [101] used a MATLAB toolbox for the design, construction and validation of a serial GB models of a CSTR. ANN model was used for the estimation of parameters of the WB model. Barkman [102] developed a serial GB model to estimate concentration distributions in the context of modeling a reaction-advection-diffusion system and was evaluated on a one-dimensional and a two-dimensional instance of the reaction system. ANN was used to estimate the parameter of the reaction model.

5.9.2. Heat Treatment Processes

Pearson et al. [103] devised a serial GB model approach for three classes of block-oriented models: Wiener models, Hammerstein models, and the feedback block-oriented models. The approach was illustrated for prediction and control of heat release inside the reactor distillation column where the least square method was used to estimate parameters of the model. Weyer et al. [104] used a serial GB model for fault diagnosis, i.e., settled material breaking away from the heat transfer surface, of a heat exchanger. The recursive least-squares identification method was used as BB model. Miao et al. [105] developed a serial GB model for the prediction of outlet temperatures of plate heat exchangers. A parameter identification method was established through the Taylor series. Cubillos et al. [106] developed a serial GB model to estimate heat losses and control of the
temperature of the combustion chamber of a pilot-scale vibrating fluidized dryer. ANN was used for parameters estimation. Farooq et al. [107] developed a serial GB model to predict the temperature of stratified virtual layers in a boiler. The nonlinear least-squares optimization method and trust-region reflective algorithm were used for the estimation of the WB model’s parameters. Aprile et al. [108] devised a serial GB model for prediction of gas utilization efficiency and the heating capacity of a water-source gas-driven absorption heat pump. Linear interpolation was used for parameter identification. Sossan et al. [109] developed a serial GB model for predictive control (MPC) of electricity consumption of a refrigeration system. Stochastic differential equations (SDEs) estimated by maximum likelihood estimation (MLE) was used as a BB model. Petersen et al. [110] investigated the use of a serial GB modeling for prediction of the residual moisture, the temperature, and the particle size in each stage for complete drying process in a multi-stage spray dryer. The WB model’s parameters were identified using the least-squares method. De-Moor et al. [111] used a serial GB strategy for the prediction of mass concentration and temperature within an imperfectly mixed fluid. The total least square method was used to identify parameters of the WB model.

5.10. Application Summary

Extracts from the reviewed papers are summarized in Tables 1–3. The percentage distribution of GB models in terms of the type of process industries is shown in Figure 2. The chemical, biochemical and pharmaceutical industry collectively have a share of 20% followed by iron and steel at 11%, oil and gas processing at 9%, materials process and energy materials at 7%, industrial robots at 7%, food industry at 5%, power plants at 5%, and water treatment at 5%. The miscellaneous studies have a collective share of 31% with further breakdown into heat transfer systems at 57% followed by reactive systems at 47%. The percentage share of types of GB sensors, i.e., parallel, serial and combined, is shown in Figure 3. The serial GB models get a share of 84% followed by parallel GB at 11% and combined GB models at 2%. Of all the cases, 3% used all three types of GB models. The percentage share of BB methods used in the GB design is plotted in Figure 4. ANN and it’s variants have the highest share of 30% followed by LS methods at 26%, LS-SVM at 7%, GA at 6%, EKF at 4%, Taylor series at 4%, and fuzzy inference at 4%; other methods collectively get a share of 19% however their individual share is 1% or lesser. Percentage in terms of types of application is shown in Figure 5. Estimation (only) got the highest share of 59% followed by estimation and control at 35%, fault diagnosis at 4% and estimation and optimization at 2%. 
Table 1. GB models application in iron and steel, food processing, chemical, biochemical, and pharmaceutical industries.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Industry</th>
<th>Application Category</th>
<th>Process</th>
<th>GB Type</th>
<th>Target</th>
<th>BB Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>[32–36,92]</td>
<td>Iron and steelmaking</td>
<td>estimation and control</td>
<td>&quot;pickling process&quot;, &quot;continuous casting&quot;, &quot;hot strip mill&quot;</td>
<td>&quot;serial&quot;, &quot;parallel&quot;, &quot;combined&quot;</td>
<td>&quot;concentration of hydrochloric acid&quot;, &quot;tundish temperature&quot;, &quot;scale breaker entry temperature&quot;, &quot;drying rate&quot;</td>
<td>&quot;Taylor series&quot;, &quot;PLS and RF&quot;, &quot;ANN&quot;</td>
</tr>
<tr>
<td>[37–39]</td>
<td>Food industry</td>
<td>estimation and control</td>
<td>&quot;fish drying process&quot;, &quot;milk drying process&quot;, &quot;whey separation&quot;</td>
<td>serial</td>
<td>&quot;drying rate&quot;, &quot;moisture contents&quot;, &quot;membrane fouling&quot;</td>
<td>&quot;ANN&quot;, &quot;exponential static membrane resistance function&quot;</td>
</tr>
</tbody>
</table>
Table 2. GB modeling application in power plant, oil and gas industry, waste water treatment, and material process and energy materials.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Industry</th>
<th>Application Category</th>
<th>Process</th>
<th>GB Type</th>
<th>Target</th>
<th>BB Type</th>
</tr>
</thead>
</table>

Table 3. GB modeling application in industrial robots and miscellaneous cases.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Industry</th>
<th>Application Category</th>
<th>Process</th>
<th>GB Type</th>
<th>Target</th>
<th>BB Type</th>
</tr>
</thead>
</table>
Figure 2. GB models in process industries in term of percentage share; (a) collective, (b) miscellaneous.

Figure 3. Percentage share of types of GB sensors, i.e., parallel, serial and combined.
5.11. Prospects and Challenges in Industry 4.0

The data gathered in industrial processes will exponentially increase with the emergence of IoT devices. The big data gives a challenge in the form of ‘Four Vs’: volume, velocity, variety, and veracity. The huge volume of datasets may require storing and processing capacity [23]. The speed through which data is collected by the integrated smart sensors of industry 4.0 will need infrastructural change. In addition to volume and velocity, the cyber manufacturing could generate variety of data based on monitoring different parts of the system, measure different phases of the process, which can be sampled at very different frequencies. The undesired samples collected due to the highly speedy sensors add veracity to the database. This type of data is not representative of the process and create an extremely heterogeneous data structure. In this context, outliers, missing data, noises, delays and data synchronism will need to be addressed [22]. To use the big data in process inference, i.e., monitoring, control, optimization, data processing methods need to be applied in the initial phase. These methods include data wrangling, visualization, sparsity and regularization, optimization, reducing dimensionality, measuring distance, representation learning, and sequential learning [24–27].
With the emergence of big data in the process industry, the GB model will need to be equipped for effectively dealing with the new scenario. The WB modeling part of the GB will mostly remain the same; however, the data-driven part will be affected by the massive amount of process data. In this context, several studies have been reported on soft sensor design based on big data, i.e., taken from IoT sensors, [17–21]. Although these studies were based on BB soft sensors, they are summarized here to understand the additional tasks in developing the BB part of the GB soft sensors in the context of the big data.

Klusch et al. [20] developed an IoT sensing system for a hydraulic aggregate consisting of an oil tank and electric motor pump. Eighteen sensors were used to monitor physical parameters such as pressure, air and oil temperature, and vibration, etc., of the oil pump in the aggregate. A stream of 50,000 data samples from 18 sensors per minute was collected. Feature reduction and annotation were performed on the basis of statistic and semantic components of the system. Then, an integration of statistical, probabilistic and semantic data analysis was used for fault detection and diagnosis of the system.

He et al. [18] developed an IoT-enabled manufacturing technology testbed (MTT) system based on temperature sensors. The sensing system comprised of 28 IoT temperature sensors attached to a CSTR plus corresponding data acquisition, transmission and storage systems. The IoT-enabled MTT made it possible to measure the real temperature distribution without assuming ideal mixing. It was observed that the IoT sensors exhibit noisy or spiky behavior at the steady state. In addition, the sensor readings fall at fixed grids and most of the IoT sensors show different levels of persistent bias. A variation in sample collection interval was also noted. Then, a statistics pattern analysis (SPA) was developed to deal with the big data related issues and effectively perform fault detection and diagnosis of the system.

Shah et al. [17] developed an IoT testbed for a multi-stage centrifugal pumping system. Non-invasive IoT vibration sensors were attached to centrifugal pumping system. The veracity of the data, i.e., unequal sampling intervals, significant noise and missing values, and its impact on data analytics was investigated. It was found that the use of Lomb's algorithm can effectively handle the data veracity. Furthermore, they devised a method of dealing with the challenge of volume and velocity. Finally, a framework of process monitoring based on data-driven predictive models for flow rate inside the pipe and speed of the pump motor was devised.

Syafrudin et al. [19] attached IoT-based sensors to the desk of a workstation in the assembly line to sense temperature, accelerometer, humidity, and gyroscope sensors. The massive data collected through the sensors were saved in the MongoDB database. An outlier detection approach was devised using the clustering technique. Then data-based fault classification of the assembly line was developed. In addition, a history of the temperature, accelerometer, humidity, and gyroscope data were displayed to the manager in real-time via a web-based monitoring system.

6. Conclusions

GB models are developed through the integration of WB and BB models. GB models have been getting the attention of researchers due to their higher intuitiveness than the BB models and high estimation accuracy than the standalone WB models. GB models are further classified into three categories namely parallel, serial and combined GB models. In the parallel GB models, BB models are used to compensate the error of WB models of the process. In the serial GB models, BB models are used to estimate parameters of the WB model. The combined GB sensors integrate the parallel and serial GB models to realize higher prediction accuracy. Applications of GB models in the process industry have been reported in iron and steel making, food processing, power plants, chemical, biochemical, pharmaceutical, water treatment, oil and gas processing, material processing, energy materials, and industrial robot.

Chemical, biochemical and pharmaceutical industry have a collective share of 20% followed by iron and steel at 11%, oil and gas processing at 9%, materials process and energy materials at
In terms of percentage share of types of GB sensors, the serial GB models got a share of 84% followed by parallel GB at 11% and combined GB models at 2%. Of all cases, 3% used all three types of GB models. In terms of percentage share of BB methods used in the GB design, ANN and its variants have the highest percentage share of 30% followed by LS methods at 26%, LS-SVM at 7%, GA at 6%, EKF at 4%, Taylor series at 4%, and fuzzy inference at 4%. In terms of percentage of types of application, estimation (only) got the highest share of 59% followed by estimation and control at 35%, fault diagnosis at 4%, estimation and optimization at 2%. 

To use the big data in process inference, i.e., monitoring, control, optimization, data processing methods need to be applied in the initial phase before development of data-based or GB models. The data pre-processing methods include data wrangling, visualization, sparsity and regularization, optimization, reducing dimensionality, measuring distance, representation learning, and sequential learning. These data preprocessing techniques are required in realizing highly efficient BB models. However, GB models, being dependent on BB models, will also need these techniques if a GB model is to be applied to Industry 4.0.

**Author Contributions:** Conceptualization, I.A.; methodology, I.A.; formal analysis, I.A.; investigation, I.A.; A.A.; writing—original draft preparation, I.A.; A.A.; writing—review and editing, I.I.C.; M.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The authors declare no conflict of interest.

**Abbreviations**

The following abbreviations are used in this manuscript:

- ALAMO: Automatic learning of algebraic models
- ANN: Artificial neural network
- ARMAX: Autoregressive moving average exogenous
- ARXIs: Auto-regressive with exogenous
- BB: Black-box
- CDU: Crude distillation unit
- CSTR: Continuous stirred tank reactor
- DR: Data reconciliation
- EKF: Extended Kalman Filter
- FIS: Fuzzy inference system
- GA: Genetic algorithm
- GB: Gray-box
- IoT: Internet of things
- LS: Least squares
- LS-SVM: Least squares support vector machine
- MF: Membership function
- MLE: Maximum likelihood estimation
- MLSSVM: Multi-I/O least squares support vector machine
- MPC: Model for predictive control
- MTT: Manufacturing technology testbed
- MWD: Molecular weight distribution
- OPFNN: Orthogonal polynomial feed forward neural network
- PCA: Principle component analysis
- PLS: Partial least squares
- PV: Photovoltaic
- RF: Random forests
- RNN: Recurrent neural network
RTO Real time optimization
SB Scale breaker
SCR Selective catalytic reduction
SDEs Stochastic differential equations
SDO Simulink design optimization tool
SMR Steam methane reforming
SPA Statistics pattern analysis
SPSA Simultaneous perturbation stochastic approximation
SSM State space model
SVM Support vector machine
TSE Twin screw extruder
WB White-box
WLLS Weighted logarithmic least squares
WRRFs Water resource recovery facilities

References


67. Wernholt, E.; Moberg, S. Nonlinear gray-box identification using local models applied to industrial robots. Automatica 2011, 47, 650–660. [CrossRef]


89. Tulleken, H.J. Grey-box modelling and identification using physical knowledge and Bayesian techniques. *Automatica* 1993, 29, 285–308. [CrossRef]


© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).