Titre: Title:	Fusion of heterogeneous industrial data using polygon generation & deep learning
Auteurs: Authors:	Mohamed Elhefnawy, Mohamed-Salah Ouali, Ahmed Ragab, & Mouloud Amazouz
Date:	2023
Type:	Article de revue / Article
Référence: Citation:	Elhefnawy, M., Ouali, M.-S., Ragab, A., & Amazouz, M. (2023). Fusion of heterogeneous industrial data using polygon generation & deep learning. Results in Engineering, 19, 11 pages. https://doi.org/10.1016/j.rineng.2023.101234

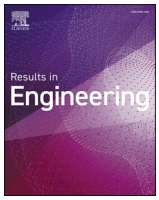
Titre de la revue: Journal Title:	Results in Engineering (vol. 19)
Maison d'édition: Publisher:	Elsevier B.V.
URL officiel: Official URL:	https://doi.org/10.1016/j.rineng.2023.101234
Mention légale: Legal notice:	©2023 Crown Copyright and The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Results in Engineering 19 (2023) 101234

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering



Fusion of heterogeneous industrial data using polygon generation & deep learning

Mohamed Elhefnawy [a,b], Mohamed-Salah Ouali [a], Ahmed Ragab [a,b,c,*], Mouloud Amazouz [b]

[a] Department of Mathematics and Industrial Engineering, Polytechnique Montréal, 2500 Chemin de Polytechnique, Montréal, Québec, H3T 1J4, Canada
[b] CanmetENERGY-Natural Resources Canada, 1615 Lionel Boulet Blvd., P.O. Box 4800, Varennes, Québec, J3X 1P7, Canada
[c] Faculty of Electronic Engineering, Menoufia University, 32952, Menouf, Egypt

ARTICLE INFO

Keywords:
Data fusion
Decision fusion
Deep learning
Polygon generation
Energy efficiency
Process system engineering

ABSTRACT

Analysis of industrial data imposes several challenges. These data are acquired from heterogeneous sources such as sensors, cameras, IoT, etc, and are stored in different structures and formats with different sampling frequencies. They are also stored in isolated silos in different locations which hinders their exploitation. Therefore, there is a clear need to integrate these disconnected data silos at different processing levels and make them clean, easily accessible, and fully exploitable. This paper proposes a data fusion method that merges heterogeneous sources of data at raw, information, and decision levels using polygon generation and deep learning (DL) techniques. An innovative polygon generation technique is proposed to preprocess each data source and convert it into powerful representations that capture all possible relationships in the data, thus extracting the maximum knowledge and achieving better prediction accuracy of the corresponding DL method. The proposed method is targeting challenging data modeling problems found in industrial processes. It is validated successfully using a case study in the realm of process system engineering. The results obtained demonstrate that the proposed fusion method is more accurate, with a minimum of 20% improvement, compared to other methods previously used in the literature.

1. Introduction

Many industries have recently undertaken a series of digitalization projects, aiming at automating and improving the operation of their processes. As a result, huge volumes of data, referred to as *Massive* or *Big Data*, consisting of a large number of observations, are being acquired from the industrial processes at a high velocity and in a variety of forms [1]. Process industries such as Pulp & Paper production, Oil refining, cement manufacturing, chemicals manufacturing, mining, and metal processing represent a significant share in terms of energy consumption as well as economic and environmental impacts [2]. These industries started investing in data analysis for better monitoring and maintenance of their equipment, units, and processes. As an example, in the process industry, pulp and paper mills started leveraging data to monitor the state of papermaking machines [3–5].

Industrial plants are equipped with hundreds or thousands of sensors (temperature, pressure, vibration, etc.) that produce large amounts of data. The analysis of these industrial data imposes several challenges, briefly stated in the following points: 1) they are acquired from heterogeneous sources such as sensors, cameras, IoT, etc., and are stored in different formats, 2) the storage of these data in isolated silos located in different locations, 3) data quality challenges due to missing values, contamination with different sources of noise, different sampling frequencies, etc. need to be addressed to derive meaningful insights, and 3) selecting between many data management platforms and communication protocols is challenging. All these challenges hinder the exploitation of these data that is an important asset.

It is too difficult to capture the whole picture of a large-scale industrial process without using multiple data sources and types. Every data type gives a complementary insight into the operation of a process. Accordingly, it is important to maximize the value of each data source through an efficient data modeling and fusion approach. The full exploitation of such data can significantly impact the performance and robustness of the developed digital solution regardless of the adopted modeling approach. The question is how industry can fuse disconnected data silos and make them clean, integrated, easily accessible, and fully exploitable?

* Corresponding author. Department of Mathematics and Industrial Engineering, Polytechnique Montréal, Canada.
E-mail addresses: mohamed.elhefnawy@polymtl.ca (M. Elhefnawy), mohamed-salah.ouali@polymtl.ca (M.-S. Ouali), ahmed.ragab@polymtl.ca, ahmed.ragab@canada.ca (A. Ragab), mouloud.amazouz@canada.ca (M. Amazouz).

https://doi.org/10.1016/j.rineng.2023.101234
Received 30 October 2022; Received in revised form 26 April 2023; Accepted 2 June 2023
Available online 18 June 2023

Nomenclature

X_j	The j^{th} variable in the data
$\widehat{q}$	The unit vector in the x direction
$\widehat{l}$	The unit vector in the y direction
$\overline{X_j}$	The mean of variable X_j
δ_j	The standard deviation of variable X_j
x_{hj}	The value of variable X_j for h^{th} observation
Z_{hj}	The standardized value of variable X_j for h^{th} observation
$\widehat{X_j}$	The unit vector of the polygon side representing variable X_j
$\overrightarrow{X_j}$	The point coordinates of the zero standardized value of the variable X_j
$\overrightarrow{X_j^h}$	The point coordinates of the standardized values for h^{th} observation of variable X_j

Most existing data integration and fusion methods assume impractical assumptions about the distribution of the industrial data [6]. They still rely on the human expert with significant knowledge about the industrial process operation. However, it is hard for an expert to efficiently address a complex system operation given its high non-linear dynamics and non-stationarity due to elevated interactions between several components. To extract useful knowledge from heterogeneous data sources with minimal human effort, artificial intelligence (AI) algorithms can play a significant role to automate the data fusion process and to develop accurate prediction models representing complex interactions and phenomena in the process thus resulting in better monitoring, optimization and maintenance [7].

Deep learning (DL) has been used as an efficient AI prediction approach (Yoshua Bengio, 2017). However, most of existing DL methods focus on developing more specialized algorithms rather than focusing on improving the quality of data and testing its representativeness before the training process. In data-centric AI, the goal is to develop AI models that can learn from quality data and produce accurate, reliable, and useful insights and predictions [8]. Data needs to be relevant, and representative of the problem being addressed before feeding it into the DL training process [9,10]. An efficient data representation method based on *Polygon Generation* was proposed in Ref. [5], where the numerical data is systematically transformed into fully representative graphs (polygons). These polygons are fed, in the form of 2-D images, as an input to train DL architectures that can achieve a great performance in computer vision problems for data classification and regression [11].

This paper proposes a novel fusion method that merges heterogeneous data sources at different abstraction levels (raw, information and knowledge levels) through polygon generation and the DL methods to enhance the industrial data representation/quality and modeling performance. The proposed method is generic and can be applied to several modeling problems having diversified data sources. The proposed method uses two fusion approaches where the main block in each of them is the polygon generation. The first fusion approach merges the knowledge acquired from different trained DL prediction models. The second one concatenates raw data and/or information then uses the concatenated data/information to build a single DL fusion model. The comparison between both approaches is discussed in this paper. The PRONTO benchmark dataset, introduced in Ref. [12], comprising diversified data sources (e.g. ultrasonic sensors, high frequency pressure sensors, process data, etc.) is used as a case study to validate the proposed fusion method.

This rest of the paper is structured as follows: Section 2 presents the background and work related to polygon generation. Section 3 presents the proposed data fusion method along with its two approaches. Section 4 presents the case study including a process overview, experimental setup, and results. Section 5 gives remarks and refers to future research directions, and Section 6 concludes the paper.

2. Background & related work

2.1. Deep learning & data quality in decision making

Digitalization and data fusion are key processes that help organizations ultimately reach well-informed strategic decisions [13]. The DIKW Pyramid, an acronym for Data, Information, Knowledge, and Wisdom, is a foundational framework used to explain how data can be transformed into actionable insights [14]. The DIKW pyramid serves for understanding the complex relationship between data fusion and digitalization to empower industries to harness the full potential of their data, leading to increased operational efficiency, and the creation of value through their value chains. The raw data is the base of the pyramid and is collected from different sources. The next level is the information that is extracted from this raw data in the form of features, and then comes the knowledge discovered in the form of ML trained models for the sake of human decision making (wisdom).

It is worth mentioning that the majority of existing data fusion methods focus only on merging data at the information and knowledge levels. This can result in a loss of information content in the original raw data. Accordingly, incorporating the data fusion at the raw level would maximize the global value of the heterogeneous data and better leverage its knowledge content thus help addressing different decision-making problems.

The adoption of AI in process industries is linked with the success of the entire AI lifecycle and addressing its needs [15]. One of the main components of this lifecycle is data preparation and representation. Given the available high performance computing (HPC) infrastructures, deep learning (DL) has been used extensively as an efficient end-to-end AI approach for building more accurate and representative models compared to other classical analytical methods [16–19] and developing robust models for several applications including system performance prediction [20]. In fact, practitioners and researchers in the DL field are still developing new architectures and/or optimizing the existing ones without looking over the quality of available data and maximize their value before exploitation. Since data acts as the fuel for the DL architectures in the modeling process, the quality of the prepared data significantly impacts the overall performance of the trained models [21]. Data-centric AI is emerging approach that aims at improving the data quality through an efficient representation to maximally exploit the DL modeling capability, thus achieving more accurate prediction [22,23]. Researchers and practitioners are recently starting using this approach to obtain the best data representation that achieve the highest prediction performance using the same DL architecture [24–26].

A number of review papers and comprehensive surveys compiling various DL-based fusion methods have been identified in the literature [27–31]. Table 1 lists examples of recent DL-based fusion methods. One of the key limitations of the existing methods is their inadequate representation of the available data, leading to suboptimal exploitation of its informational content. Moreover, these methods focus on the fusion at feature (information) or decision (knowledge) level, potentially causing information loss compared to using the raw data. To address these shortcomings, this paper proposes a fusion method that merges heterogeneous data sources across three levels (raw, information and knowledge) while using an effective data representation technique to maximize data exploitation. This approach is detailed in the following subsection.

2.2. Polygon generation: an efficient data representation

The *Polygon Generation* was proposed in Ref. [5] as an efficient data representation method, where the numerical data is systematically

Table 1
Examples of existing DL-based data fusion methods.

Reference	Fusion level	Description	Applications
[32]	Feature (information)	Extracting features for each fault using stacked autoencoders (SAE) and merging them for accurate diagnostics	Fault diagnostics in the Tennessee Eastman process
[33]	Feature (information)	Merging spatial and temporal features using Convolutional neural networks (CNN) and Long-short term memory (LSTM) for accurate diagnostics	Fault diagnosis in industrial coking furnace process
[34]	Feature (information)	Fusing sensory data using parallel CNN for predicting the tool wear compared to other ML and DL regressors	Cutting tool monitoring and bearing fault diagnosis
[35]	Decision (knowledge)	Merging diversified remote sensing data: LiDAR, Hyperspectral data and high-resolution RGB images using ensemble of classifiers	Classification of urban land use and land cover

transformed into fully representative graphs (polygons). Each variable is represented as a polygon side where every point on the side represents a corresponding numerical value. The points on the polygon sides representing the observation values are connected in the form of the Hamiltonian cycles [36]. As a definition, Hamiltonian cycle is a close loop that can go through all vertices only once. These generated polygons can represent all complex interrelationships between data variables. In what follows, we summarize the steps of the polygon representation proposed in Ref. [5] through a toy example of data observations with six variables.

Each variable X_j, where $j = 1, 2, .., 6$ is represented by a polygon side, as shown in Fig. 1 (hexagon in this example). All variables are numbered in a clockwise direction. Eq. (1) is used to calculate the standardized values Z_{hj} for the variable X_j of the h^{th} observation in the data.

$$Z_{hj} = \frac{x_{hj} - \overline{X_j}}{\delta_j} \tag{1}$$

where x_{hj} is the numeric value of the h^{th} observation for the variable X_j, $\overline{X_j}$ and δ_j are the mean value and standard deviation of variable X_j, respectively. As shown in Fig. 1, the standardized values Z_{hj} are represented by the point coordinates (in orange) on each side of the polygon, calculated using Eq. (2).

$$\overrightarrow{X_j^h} = \overrightarrow{X_j} + \left(Z_{hj} * \widehat{X_j}\right) \tag{2}$$

where $\overrightarrow{X_j}$ represents the point coordinates (in blue) of the zero standardized value of the variable X_j and $\widehat{X_j}$ represents the unit vector of each corresponding polygon side. Table 2 shows the values calculated using Eq. (1) and Eq. (2) for the toy example shown in Fig. 1. A pseudocode for the polygon generation is shown in Table 3.

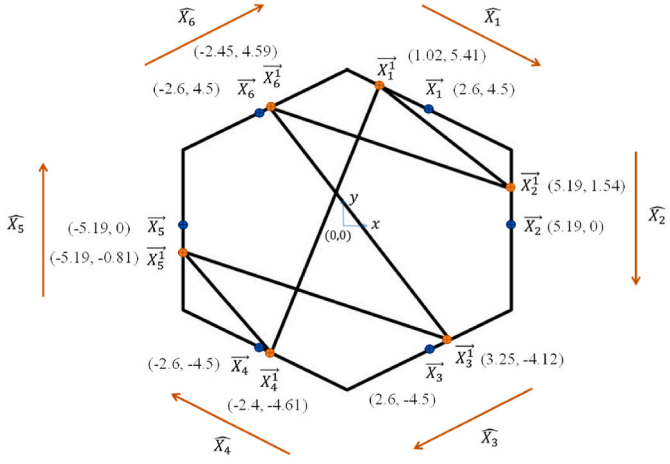


Fig. 1. A polygon generated from a numeric observation of six data variables using the method proposed in Ref. [5], where $\overrightarrow{X_j^h}$ represents the point coordinates of standardized values of the h^{th} observation for each variable X_j, $\overrightarrow{X_j}$ represents the point coordinates of the zero standardized value of the variable X_j and $\widehat{X_j}$ represents the unit vector of each polygon side.

2.3. Polygon generation for industrial modeling problems

The Polygon Generation method [5] opens the door for modeling different industrial problems from computer vision perspectives, thus exploiting the DL capability in discriminative and generative modeling for efficient mapping of the input process variables into the desired outcomes [38]. This data representational method outperformed other machine learning and state-of-art DL classifiers [5]. In that work, it is used for fault classification in each of the targeted case studies: Tennessee Eastman Process (TEP) [39] and a reboiler system of heat recovery network in a thermomechanical pulp mill. In Ref. [11], the polygon generation is used along with the state-of-art generative DL technique (conditional generative adversarial networks – cGANs) to predict the key performance indicators using a complex dataset acquired from a black liquor recovery boiler in a Kraft pulp & paper mill located in Canada. Besides, the polygon generation is used for time-series prediction in a highly dynamic non-linear industrial processes, where video-to-video translation technique is used for mapping polygon streams (videos) representing input process variables into those representing the outputs [11].

Table 4 provides an overview and categorizes industrial problems modeled using polygon generation method proposed in Ref. [5]. The common limitation of the three polygon generation methods is their reliance on a single data source or homogeneous sources for predicting categorical, continuous, or time-series outputs. This paper addresses this limitation by fusing heterogeneous sources to leverage the complementary information about the targeted complex physical phenomena. This fusion approach yields improved data quality and efficient prediction, ultimately leading to more effective decision-making process.

3. Proposed data fusion method: polygon & decision fusion

This paper proposes a data fusion method that merges different data sources at raw, information (polygon level) and decision level for more efficient and accurate modeling of the industrial processes. The proposed method comprises two different fusion approaches: The first approach (called decision fusion) merges the knowledge obtained from the different DL models, and the second one (polygon fusion) fuses the raw data and/or information in the form of a single representative polygon. The details of each approach are illustrated in the following subsections.

3.1. Approach 1: decision fusion

The first approach fuses the different data sources at the knowledge (decision) level. As shown in Fig. 2, the data sources are referred to as data blocks (DB). In case of N data blocks (DB_1, DB_2, …, DB_N), the polygon generation technique proposed in Ref. [5] is applied to each DB separately to synthesize a set of representative polygons for each data source. This means that every observation in a data block is converted into set of polygons (images) and fed as an input to train the corresponding DL model. Given N data blocks, a DL model is trained for each block, denoted as $DL\ model_1, DL\ model_2, \ldots, DL\ model_N$. For readers who

Table 2

Calculations of point coordinates $\overrightarrow{X_j^1}$ on the sides of the polygon for a numeric observation with six variables shown in Fig. 1, where $\hat{q}$ and $\hat{l}$ are the unit vectors of x and y directions, respectively.

j	x_{1j}	$\overline{X_j}$	δ_j	Z_{1j}	$\widehat{X_j}$	$\overrightarrow{X_j}$	$\overrightarrow{X_j^1}$
1	14.25	28.56	7.85	−1.82	$0.87\,\hat{q} - 0.5\,\hat{l}$	$2.6\,\hat{q} + 4.5\,\hat{l}$	$1.02\,\hat{q} + 5.41\,\hat{l}$
2	10.79	23.15	8.05	−1.54	$0\,\hat{q} - 1\,\hat{l}$	$5.19\,\hat{q} + 0\,\hat{l}$	$5.19\,\hat{q} + 1.54\,\hat{l}$
3	9.83	12.04	2.93	−0.75	$-0.87\,\hat{q} - 0.5\,\hat{l}$	$2.6\,\hat{q} - 4.5\,\hat{l}$	$3.25\,\hat{q} - 4.12\,\hat{l}$
4	15.21	15.95	3.3	−0.22	$-0.87\,\hat{q} + 0.5\,\hat{l}$	$-2.6\,\hat{q} - 4.5\,\hat{l}$	$-2.4\,\hat{q} - 4.61\,\hat{l}$
5	19.78	25.1	6.54	−0.81	$0\,\hat{q} + 1\,\hat{l}$	$-5.19\,\hat{q} + 0\,\hat{l}$	$-5.19\,\hat{q} - 0.81\,\hat{l}$
6	3135	2979.07	888.42	0.18	$0.87\,\hat{q} + 0.5\,\hat{l}$	$-2.6\,\hat{q} + 4.5\,\hat{l}$	$-2.45\,\hat{q} + 4.59\,\hat{l}$

Table 3

Pseudocode of the Polygon Generation method.

Algorithm 1: Polygon Generation

Inputs: Raw numeric data matrix X, with m rows (representing the observations) and n columns (representing variables), pre-specified side length of the regular polygon (l)

Construct a polygon for all observations

Construct a Hamiltonian cycle matrix (*HamMat*) using Hamiltonian decomposition using Lucas–Walecki Hamiltonian decompositions for N-complete graphs [37]

Do $j = 1$ *to* n

Calculate the midpoint of each polygon side ($\overrightarrow{X_j}$) and its unit vector ($\widehat{X_j}$)

Calculate the mean of each variable ($\overline{X_j}$) and its standard deviation (δ_j)

End

Do $h = 1$ *to* m

Do $j = 1$ *to* n

Calculate the standardized data values Z_{hj} using Eq. (1)

Calculate the data coordinates on the polygon sides $\overrightarrow{X_j^h}$ using Eq. (2)

End

Connect the points $\overrightarrow{X_j^h}$ on the polygon sides using *HamMat* to generate $float(n/2)$ polygon images for each h^{th} observation

End

Outputs: Representative polygons that systematically express the interrelationships between all data variables

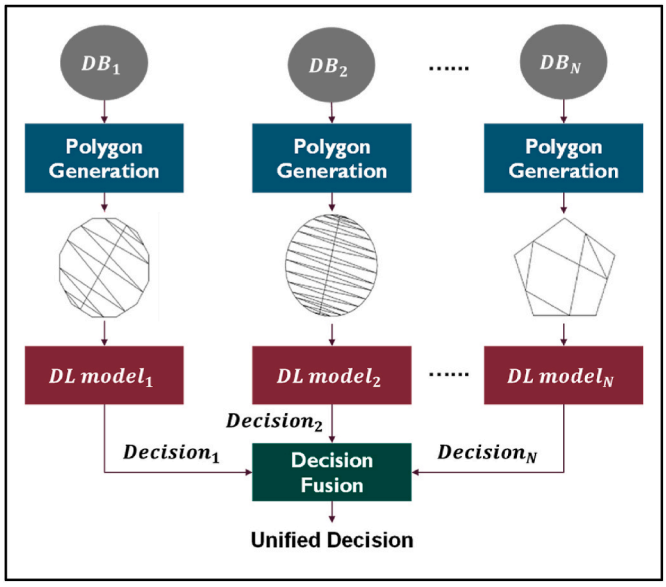


Fig. 2. Schematic diagram of *Approach 1 (Decision Fusion)*.

Table 4

Categorization of Polygon Generation-based modeling problems.

Reference	Problem Type	Description	Applications
[5]	Classification	Efficient representation of datasets with large number of variables, thus building accurate prediction models that outperform classical ML and state-of-art DL classifiers	• Fault classification in the Tennessee Eastman process • Identification of causes of an abnormal event in a reboiler system of heat recovery network in thermomechanical pulp mills
[11]	Regression	Mapping between the distributions of input variables and numerical outputs using an efficient Image-to-image DL translation technique (cGAN)	Prediction of key performance indicators (steam production divided by black liquor flow, emitted amount of sulphur dioxide & emitted amount of total reduced sulphide) of a black liquor recovery boiler in a Kraft pulp & paper mill
[11]	Time-series Prediction	Modeling the dynamic behavior of industrial processes using CycleGAN for mapping input variables into the time-series outputs	Time-series modeling of key performance indicators (evaporated water, concentrator efficiency & fouling index) of a concentrator equipment in a complex heat exchanger network in a pulp mill

are interested in detailed procedure of the polygon generation technique, they can refer to the method described in Ref. [5].

One of the DL architectures that can deal efficiently with these polygon images is the convolutional neural networks (CNNs) [40]. The proposed approach allows the flexibility to select the neural network architectures used for building the DL models.

Each set of polygon images for each data block has a different structure according to the nature (i.e. sampling frequency, number of variables and their interactions, etc.) of the corresponding data block. Accordingly, the network architecture and its hyperparameters for each DL model need to be tuned and optimized for the best prediction performance. There are several algorithms that can be used for hyperparameter optimization in DL architectures [41]. The proposed approach also allows for selecting the neural architecture using strategies such as neural architecture search (NAS) [42]. It is worth mentioning that both hyperparameter optimization and NAS are time consuming and computationally expensive tasks, however, they provide more adaptability and reliability for the proposed fusion method.

After terminating the training of the N DL models, the unseen observations (testing data) from different DBs are converted into polygons as well and tested using the trained models to acquire N decisions denoted as $Decision_1, Decision_2, \ldots, Decision_N$. These decisions are then fused using one of different voting criteria such as majority voting, weighted majority voting, behavior knowledge space (BKS) [43], etc. This fusion approach results in a unified and accurate decision that leverages the use of different complementary data sources. The pseudocode for the proposed decision fusion approach is shown in Table 5.

Table 5
Pseudocode for the *Approach 1 (Decision Fusion)*.

Algorithm 2: Approach 1 (Decision fusion)
Inputs: N heterogeneous training data blocks ($TrainDB_1, TrainDB_2, .., TrainDB_N$), N DL models ($DL\ model_1, DL\ model_2, ..DL\ model_N$) (could be the same architecture such as CNN), N heterogeneous testing (unseen) data blocks ($TestDB_1, TestDB_2, .., TestDB_N$), Decision Fusion strategy (e.g. Majority Voting, Weighted Majority Voting, BKS) (*DF strategy*)
/***Training Phase***/
Do $i = 1$ *to* N
Convert $TrainDB_i$ to a set of $TrainPolygons_i$ using *Algorithm 1: Polygon Generation*
Feed $TrainPolygons_i$ into $DL\ model_i$ for training
End
/***Testing Phase***/
Do $k = 1$ *to* N
Convert $TestDB_k$ to a set of $TestPolygons_k$ using *Algorithm 1: Polygon Generation*
Test $DL\ model_k$ using $TestPolygons_k$ to come up with $Decisions_k$
End
Fuse $Decisions_1, Decisions_2, Decisions_N$ using *DF strategy* to obtain a **Unified Decision**.
Outputs: Trained DL models ($DL\ model_1, DL\ model_2, ..DL\ model_N$) for predicting new (unseen) observations.

3.2. Approach 2: polygon fusion

One of the strengths of the polygon generation proposed in Ref. [5] is that it can deal with numerical data with a large number of variables. The second approach makes use of this advantage and fuses the raw data and/or information at the polygon level. In this case, each DB can be the raw data, or features extracted from the original variables of the corresponding data source. A unified polygon generation is followed as shown in Fig. 3, where the DBs are concatenated at the raw and/or information (feature) level.

This approach allows synthesizing a unified set of polygons that represent all the DBs with interrelationships between the different DBs and the variables (features) within each DB. This set of polygons is then introduced as an input to train a single DL model. This trained model is then used for testing new (unseen) observations representing the different DBs to have an accurate unified decision. It is worth mentioning that the hyperparameter optimization of the selected neural network architecture in this approach is less time-consuming than that of Approach 1 (Decision Fusion), as we have a single model to tune instead of multiple ones. The pseudocode of the polygon fusion approach is shown in Table 6.

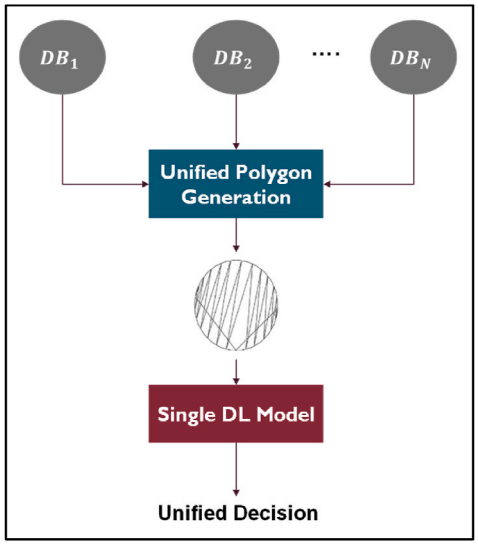


Fig. 3. Schematic diagram of *Approach 2 (Polygon Fusion)*.

Table 6
Pseudocode for *Approach 2 (Polygon Fusion)*.

Algorithm 3: Approach 2 (Polygon fusion)
Inputs: N heterogeneous training data blocks ($TrainDB_1, TrainDB_2, .., TrainDB_N$), A DL classification architecture ($DL\ model$) for training all the data blocks, N heterogeneous testing (unseen) data blocks ($TestDB_1, TestDB_2, .., TestDB_N$)
/***Training Phase***/
Concatenate ($TrainDB_1, TrainDB_2, .., TrainDB_N$) at the raw or information level to come up with a single *TrainDB*
Convert *TrainDB* to a set of *TrainPolygons* using Algorithm 1: Polygon Generation
Feed *TrainPolygons* into *DL model* for training
/***Testing Phase***/
Concatenate ($TestDB_1, TestDB_2, .., TestDB_N$) at the raw or information level to come up with a single *TestDB*
Convert *TestDB* to a set of *TestPolygons* using Algorithm 1: Polygon Generation
Test *DL model* using *TestPolygons* to come up with a **Unified Decision**
Outputs: Trained DL model (*DL model*) for predicting unseen observations.

4. Case study: multiphase flow facility (PRONTO benchmark dataset)

The proposed fusion method is validated using the benchmark dataset PRONTO (PROcess NeTwork Optimization) collected from a multiphase flow facility in the process system engineering laboratory at Cranfield University [12]. This benchmark dataset is publicly available and has been used in a number of research studies in the field of process system engineering [44–47]. The process description and data collected are presented in this section along with experimental setup to demonstrate the effectiveness of the proposed method.

4.1. PRONTO: process overview

The PRONTO dataset includes heterogeneous data collected from different sensors installed in the multiphase flow facility shown in Fig. 4 [48]. The dark red rectangles in the figure refer to the process variables, the green circles refer to high-frequency pressure measurements and the purple oval highlights the ultrasonic sensor.

In this multiphase flow facility, water and air are mixed through the horizontal section, then they are separated. The flow rates of the input air and water are controlled for implementing different operating conditions. The data are collected from these operating conditions including normal and faulty states. Input flows are mixed in the mixing zone then at the top of the riser, the flow is separated by two separators in sequence; the water returns to the storage tank and the airflow is released into the atmosphere after separation. For more detail on the process description and its operating conditions, the readers may refer to this technical report [12].

4.2. Data sources and feature extraction

Table 7 shows the heterogeneous data sources in the PRONTO benchmark dataset along with their sampling frequencies and availability. In this paper, three measured variables from different data sources are used, namely: process variables (17 variables including pressure, flow rate, temperature, and water density), nine pressure sensors distributed along the pipelines from the mixing zone to the riser top as shown in Fig. 4 and the last variable is measured by the Doppler ultrasonic sensor. It is worth mentioning that the videos, alarm, event, and change logs can be used in the future work where the polygon generation can be adapted to handle these types of data, but this is out of the scope of this work.

The main challenge in the heterogeneity of these three different sources is the variability of the sampling frequencies and their availability. As shown in Table 7, the process data is available continuously at a lower sampling rate (1 Hz), while high-frequency pressure and ultrasonic measurements are collected in 60-sec segments at much higher sampling rates. To address this problem, one solution is to synchronize

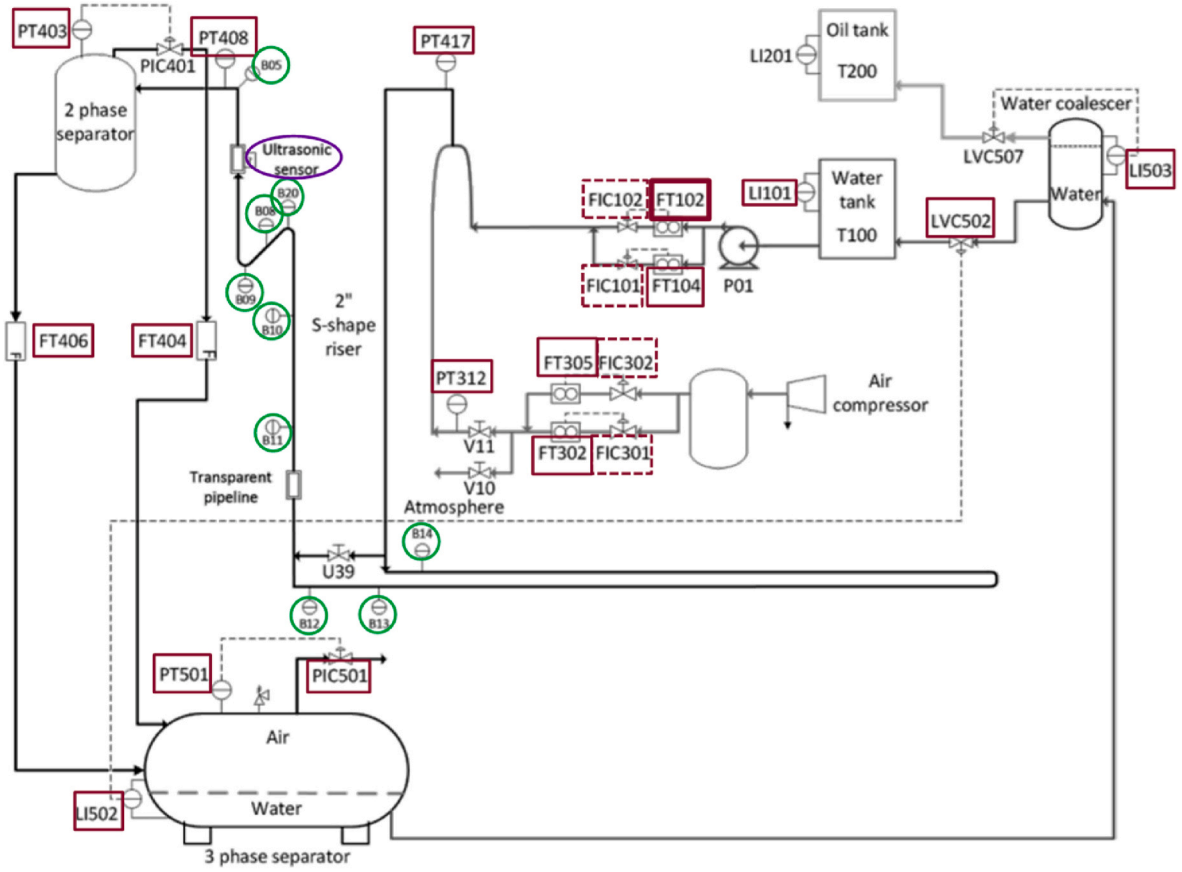


Fig. 4. A schematic diagram of the multiphase flow facility adapted from [12].

Table 7
Types of heterogeneous data in the PRONTO benchmark dataset (only data sources in bold were used in this paper).

Data sources	Sampling rate	Availability
Process variables (pressure, flow rate, temperature, and density sensors)	**1 Hz**	**Continuous**
Doppler ultrasonic sensor	**10 kHz**	**60 s**
High-frequency pressure sensors	**5 kHz**	**60 s**
Videos	–	30–60 s
Alarm, event, changelogs	Event-driven	Discrete event

the 60-s segments of process data with high-frequency measurements using the operation log, which indicates the process state (class) and corresponding timestamps.

For the sake of synchronization, a number of features are extracted for each second from the high-frequency pressure and ultrasonic data to make use of all sampled data in each data source instead of using a simple sample-and-hold strategy. Accordingly, these features are integrated with the remaining process data. There are frequency-domain representation methods that can be used for synchronizing these heterogeneous data sources such as Fourier Transform (FT) [49]. However, the limitation of the FT is that it assumes that the signal being analyzed is stationary, meaning that its statistical properties do not change over time. This is not a practical assumption in many industrial applications. Therefore, continuous wavelet transform (CWT) [50] can be used instead to reflect the non-stationarity of the data taking into account both its spectral and temporal information. It has been used extensively and successfully in the literature for fault detection and diagnosis in various industrial applications [51–53]. In this paper, five CWT features were extracted per second for each pressure sensor and the same for the ultrasonic sensor in the time-frequency domain as shown in Fig. 5. Accordingly, forty-five features were extracted from the pressure sensory data, and other five features were extracted from the ultrasonic sensory data.

4.3. Experimental setup and results

Regarding the decision fusion (Approach 1), polygons were generated for the 17 process variables and the CWT features extracted from the pressure and ultrasonic data separately as shown in Fig. 6. A CNN model was trained for each data type, then a voting criteria was used for fusing the decisions of the three different DL trained models. While in the polygon fusion approach, a unified set of polygons that represents the relationships between all sixty-seven variables and features was generated to be used for training a single CNN architecture as shown in Fig. 7. The trained model was then used as a classifier to identify normal and faulty situations.

For the purpose of validation, we compare the results obtained using the two fusion approaches with those obtained in the literature [12]. The authors in Ref. [12] have used the data of the high-frequency pressure sensors to classify the normal operation and the four different faulty scenarios (Slugging, Air blockage, Air Leakage and Diverted Flow). In that work, frequency domain features have been extracted per second for the whole 60-s measurements without overlapping. The maximum magnitude of the frequency domain spectrum for each 100 Hz band was used as a feature, accordingly, twenty-five features were extracted per second for each pressure sensor, then the Naïve Bayes algorithm has been used as a fault classifier.

In this paper, the performance of our proposed fusion method was compared with three baseline classifiers; Radial Basis Function Neural

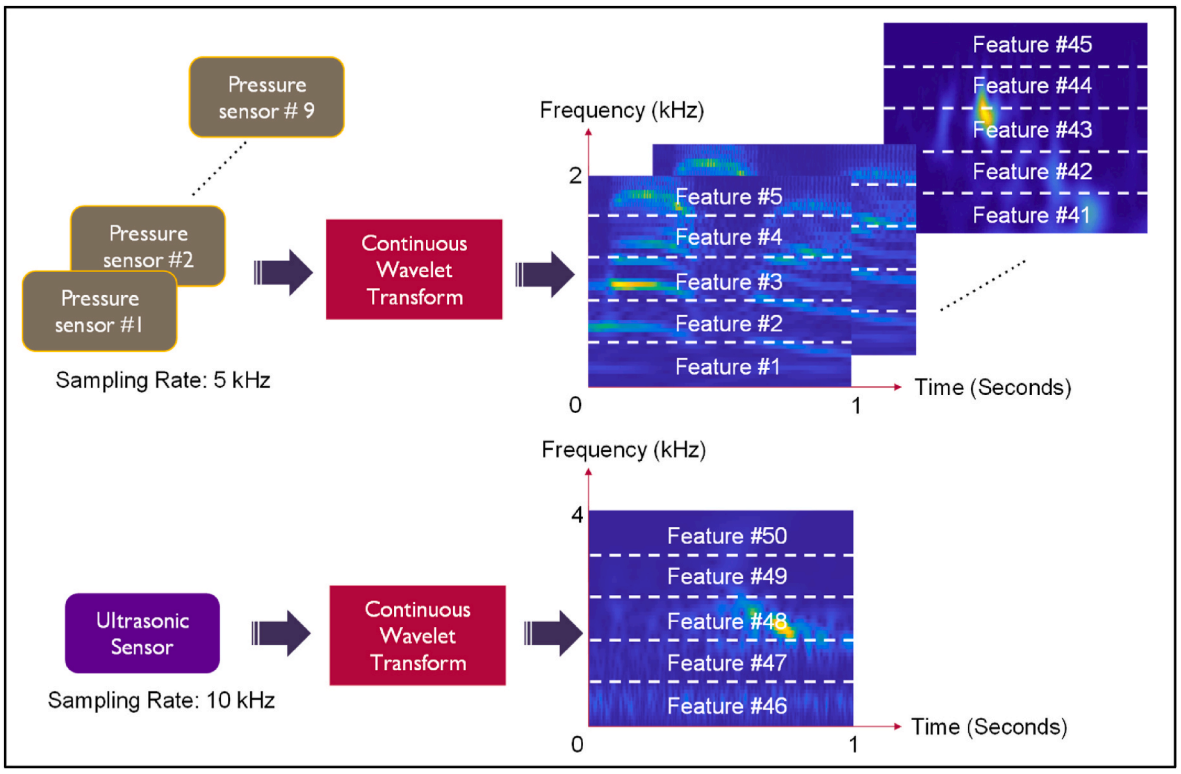


Fig. 5. Feature extraction using continuous wavelet transform for high-frequency pressure sensors and the ultrasonic sensor in the PRONTO dataset.

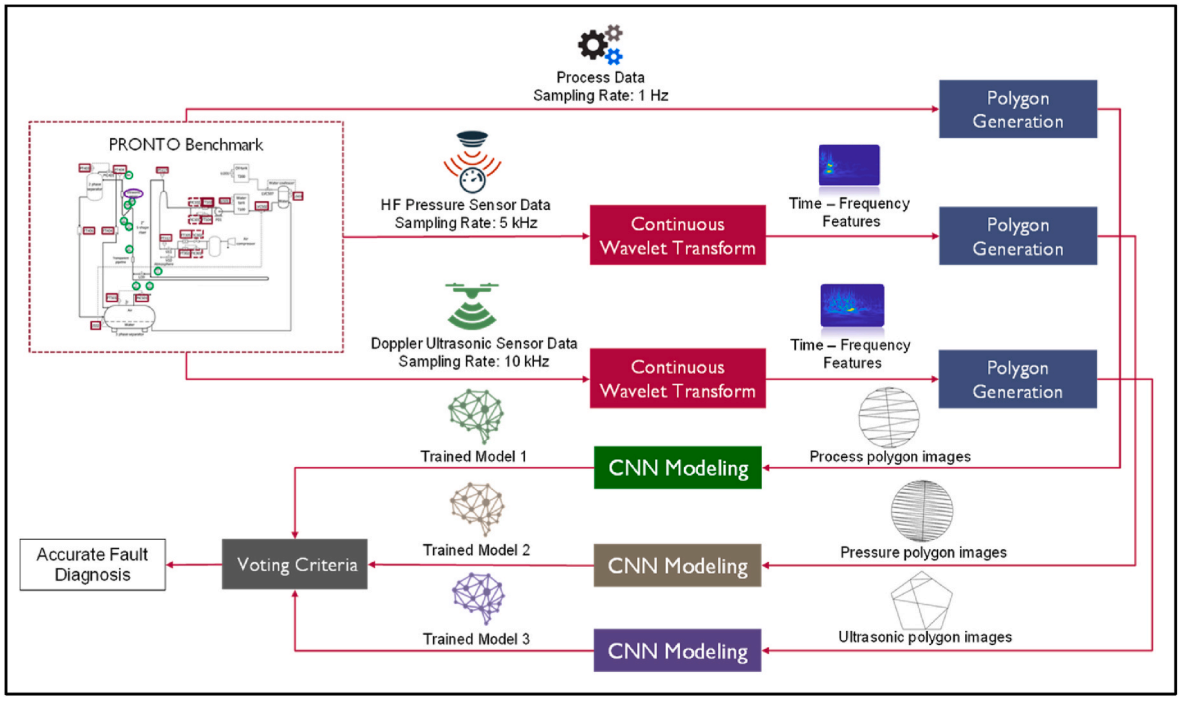


Fig. 6. Schematic of the decision fusion (Approach 1) using the PRONTO dataset.

Network (RBFNN), Multi-layer Perceptron (MLP) and k-Nearest Neighbors (kNN), given their extensive use in the literature [54–65]. The hyperparameters of all classifiers were optimized using the 5-fold cross validation technique, where the ranges of these hyperparameters are listed in Table 8. Grid search has been used to fully represent the whole space of the hyperparameters where they are optimized. This includes all possible combinations of different hyperparameters in each classifier.

The performance metrics used in this paper are the F1 score, precision, recall, and overall accuracy calculated using Eq. (3), (4), (5) & (6), respectively.

$$F1 = \frac{2\,TP}{2\,TP + FP + FN} \tag{3}$$

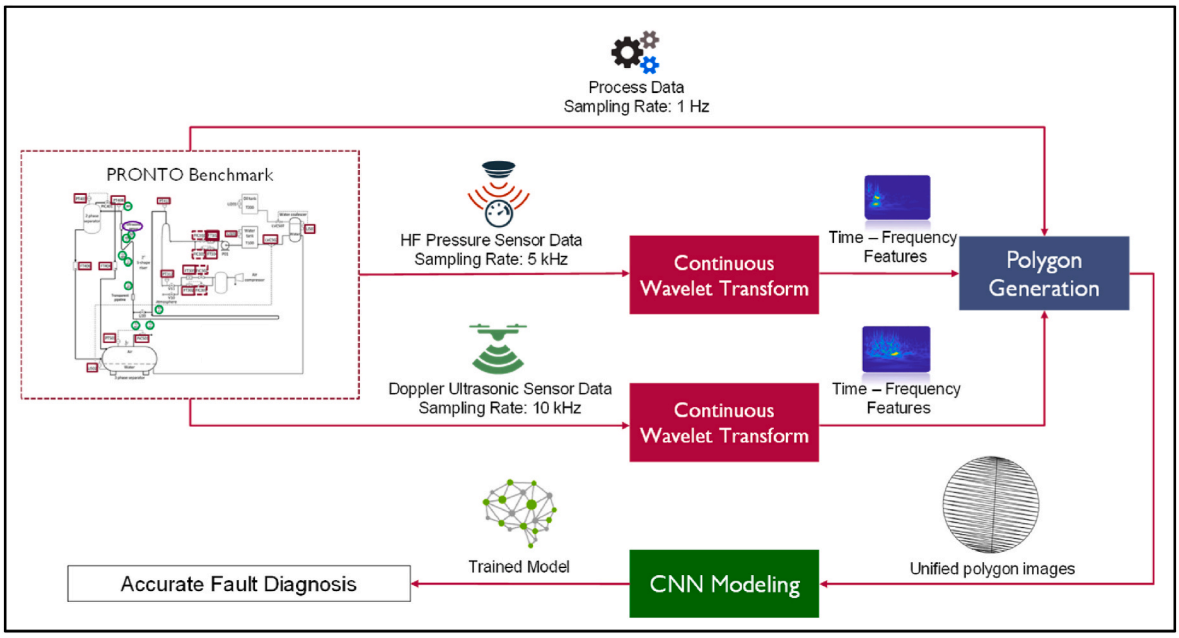


Fig. 7. Schematic of the polygon fusion (Approach 2) using the PRONTO dataset.

Table 8
Range of hyperparameters of each classifier.

Algorithm	Hyperparameters
Proposed Fusion Method	n = [4,64], r = (2,3), batch size = [50,200] # epochs = [30,150], activation function = {sigmoid, ReLU}
MLP	# neurons in hidden layer = [10,40] Maximum number of iterations = [1000,5000]
KNN	K = [3,15] Weights = {'uniform','weighted with distance'} Distance = {'Euclidean','Manhattan'}
RBFNN	# clusters = [5,30] coefficient of smoothing exponential kernel = [0.1,0.9]

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$Overall\ Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{6}$$

where *FP*, *TP* and *FN*, *TN* are the number of false positives, true positives, false negatives, and true negatives respectively. The performance metrics for each classifier are shown in Tables 9–11. The behavior knowledge space (BKS) [66] was used as the strategy of merging decisions in the proposed decision fusion (Approach 1) as a voting criteria. We partitioned the whole dataset randomly before the polygon generation and CWT feature extraction steps into 70% for training and validation for hyperparameter optimization while the remaining 30% was used for testing, as recommended by many machine learning researchers and practitioners [67–69]. The training stage is done once offline and the trained models are then saved and deployed for online classification.

The TensorFlow [70] with Python 3.7 was used to implement, train and test the proposed fusion method and other baseline algorithms using a computational infrastructure with following specifications: Intel(R) Core(TM) i7-8750H CPU @2.2 GHz + NVIDIA GeForce GTX 1070 with Max-Q Design. Accordingly, the training and testing elapsed time for each classifier are shown in Table 12.

5. Discussion, remarks & future work

It can be observed that the proposed fusion method with the two approaches outperforms the classification method used in Ref. [12] and other comparable classifiers in terms of F1 scores, precision, recall and the overall accuracy. This is mainly attributed to the fact that both fusion approaches maximally exploit the different available data sources to capture the whole picture of the process for more accurate decision-making. It can also be observed that the polygon fusion approach is slightly better than the decision fusion approach. Despite having a large number of variables and features (67 variables and features), it has been proved that the fusion at the polygon level is effective. It is worth mentioning that the elapsed time for training and testing depends on the available computational power. The training stage is done only once offline, then the testing is done as long as we have new online observations. Although the training of the DL models in our proposed method needs more time than other baselines, the testing (inference) time is relatively small which ensures the reliability of the proposed method.

The main motivation of developing the proposed fusion approaches is improving the performance of industrial systems through merging the available heterogeneous data (i.e., PRONTO dataset) by exploiting the modeling power of DL and improving the data representation, thus maximizing its global value. This is achieved through using an efficient representation approach (i.e., Polygon Generation) and predictive DL algorithms (i.e., CNNs). These two main blocks help accurately predict the system performance, which is an urgent and prioritized need for many industrial systems.

It is worth mentioning that one of the advantages of the proposed

Table 9
Overall accuracy of each classifier using the PRONTO dataset.

Algorithm	**Approach 1 Decision Fusion**	**Approach 2 Polygon Fusion**	Naïve Bayes [12]	RBFNN	MLP	*k*NN
Overall Accuracy %	**99**	**100**	75	61	69	79

Table 10
Performance metrics of each classifier using the PRONTO dataset for Normal, Slugging and Air blockage classes.

Algorithm	Normal			Slugging			Air Blockage		
	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Approach 1 Decision Fusion	**0.99**	**0.99**	**1**	**0.99**	**1**	**0.99**	**0.99**	**0.99**	**1**
Approach 2 Polygon Fusion	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
Naïve Bayes [12]	0.64	0.83	0.52	0.7	0.62	0.86	0.79	0.76	0.82
RBFNN	0.69	0.75	0.64	0.84	0.85	0.83	0.54	0.47	0.64
MLP	0.71	0.73	0.69	0.68	0.85	0.57	0.69	0.67	0.72
*k*NN	0.89	0.92	0.87	0.94	0.94	0.93	0.72	0.68	0.76

Table 11
Performance metrics of each classifier using the PRONTO dataset for Air Leakage and Diverted Flow classes.

Algorithm	Air Leakage			Diverted Flow		
	F1	Precision	Recall	F1	Precision	Recall
Approach 1 Decision Fusion	**0.99**	**1**	**0.99**	**0.99**	**0.99**	**1**
Approach 2 Polygon Fusion	**1**	**1**	**1**	**1**	**1**	**1**
Naïve Bayes [12]	0.75	0.65	0.91	0.96	0.96	0.96
RBFNN	0.66	0.78	0.58	0.42	0.43	0.41
MLP	0.64	0.75	0.57	0.71	0.62	0.84
*k*NN	0.78	0.82	0.75	0.76	0.76	0.75

Table 12
Training and testing elapsed time for each classifier using the PRONTO dataset.

Algorithm	Training elapsed time (in sec)	Testing elapsed time (in seconds)
Approach 1 Decision Fusion	300	0.3
Approach 2 Polygon Fusion	100	0.1
Naïve Bayes [12]	0.01	0.007
RBFNN	156.5	0.03
MLP	1.84	0.01
*k*NN	0.003	0.3

fusion approaches is the generic nature of the polygon generation technique and its reliability in different systems using available numerical data. Representing all interrelationships between data variables in the form of polygon images opens the door to make use of the powerful DL classifiers for computer vision such as CNN. Accordingly, the polygon generation achieves one of the data-centric AI goals of improving the industrial data quality for better DL performance. In addition, polygons can be considered as a standardized data format that facilitates the integration of multiple data sources. Besides, the polygon representation can help in encryption and information security, thus facilitating the data sharing among different departments in the plant without affecting its confidentiality. Moreover, data visualization in the form of polygons can help in interpretation of DL models in the future work. Moreover, the DL modeling is flexible, and more state-of-art architectures other than CNN can be used for data classification. In addition, it is flexible to choose voting criteria in the decision fusion process according to the targeted application and the contribution of each data source.

It is worth mentioning that the resolution of the synthesized polygon images is critical in terms of prediction performance. Increasing their resolution may significantly increase the accuracy of the whole prediction approach but will result in a more computationally expensive task. One of our future research directions is using other data sources beside the existing numerical data such as images, point cloud LiDAR data, videos, categorical data, etc. using an ensemble of DL algorithms in our fusion approaches. The goal is to build a flexible and generic big data fusion platform that provides the end-users with accurate and robust knowledge. Integrating the proposed method using big data processing tools such as PySpark [71] will allow the end users to handle heterogeneous data processing at scale through flexible deployable DL pipelines. The performance of such pipelines can be monitored and fully supervised by AI experts to ensure the reliability of the trained models.

Other heterogeneous datasets will be used for further validation of our fusion approaches. Systems other than those in the process industry such as the forestry industry can be considered using the available heterogeneous data such as LiDAR data, digital aerial photography (DAP), drone-acquired images, etc. Adaptation of the proposed fusion method to work efficiently with regression problems will be considered in our future work. The visual representation of the polygon generation may help capture the actual data distribution, thus accurately predicting continuous outputs.

6. Conclusion

This paper proposes a novel method for heterogeneous data fusion. The proposed method is implemented in two distinct approaches (decision fusion and polygon fusion), both approaches rely on polygon generation and predictive deep learning (DL) modeling. Heterogeneous data blocks are represented as set of polygon images either in the form of multiple sets in the decision fusion approach or a single set in the polygon fusion one. These polygon images are then used to train DL models with selected architectures and optimized hyperparameters for accurate prediction. The two fusion approaches are validated on a benchmark dataset in the realm of process system engineering. Despite the heterogeneity of this dataset, the proposed fusion method is successfully applied at the raw, information (feature) and decision levels. It outperformed other machine learning models used in the literature to classify the normal and different faulty scenarios in the data. This fusion method would be a cornerstone for a big data fusion platform that can help accurately predict industrial system performance. As a generic method, it will be applied to other industrial systems owing to the flexibility of selecting and optimizing the DL architectures and the use of polygon generation as an efficient data representation technique.

CRediT authorship contribution statement

Mohamed Elhefnawy: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing, Validation. **Mohamed-Salah Ouali:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Ahmed Ragab:** Conceptualization, Methodology, Supervision, Writing – review & editing, Data curation, Validation. **Mouloud Amazouz:** Writing – review & editing, Supervision, Project administration, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) under research grant number 231695 and the Natural Resources Canada's OERD (Office for Energy Research and Development) Program.

References

[1] P. Zikopoulos, C. Eaton, others, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw-Hill Osborne Media, 2011.
[2] National Inventory Report, Greenhouse gas sources and sinks in Canada canada's submission to the united nations framework convention on climate change, Executive Summary (2019). Retreived April 20, 2023 from, https://publications.gc.ca/site/eng/9.506002/publication.html.
[3] A. Ragab, S. Yacout, M.S. Ouali, H. Osman, Prognostics of multiple failure modes in rotating machinery using a pattern-based classifier and cumulative incidence functions, J. Intell. Manuf. 30 (1) (2019) 255–274.
[4] M. Nystad, L. Lindblom, Artificial Intelligence in the Pulp and Paper Industry: Current State and Future Trends, 2020.
[5] M. Elhefnawy, A. Ragab, M.-S. Ouali, Fault classification in the process industry using polygon generation and deep learning, J. Intell. Manuf. (2021).
[6] T. Meng, X. Jing, Z. Yan, W. Pedrycz, A survey on machine learning for data fusion, Inf. Fusion 57 (2) (2020) 115–129.
[7] A.M. Schweidtmann, E. Esche, A. Fischer, M. Kloft, J.U. Repke, S. Sager, A. Mitsos, Machine learning in chemical engineering: a perspective, Chem.-Ing.-Tech. 93 (12) (2021) 2029–2039.
[8] M.H. Jarrahi, A. Memariani, S. Guha, The Principles of Data-Centric AI (DCAI), 2022, pp. 1–14.
[9] N. Polyzotis, M. Zaharia, What Can Data-Centric AI Learn from Data and ML Engineering?, 2021. ArXiv Preprint ArXiv:2112.06439.
[10] E. Strickland, Andrew Ng, AI minimalist: the machine-learning pioneer says small is the new big, IEEE Spectrum 59 (4) (2022) 22–50.
[11] M. Elhefnawy, M.-S. Ouali, A. Ragab, Multi-output regression using polygon generation and conditional generative adversarial networks, Expert Syst. Appl. (2022).
[12] A. Stief, R. Tan, Y. Cao, J.R. Ottewill, PRONTO Heterogeneous Benchmark Dataset [Data Set], Zenodo, 2019.
[13] A. Chhabra, S. Williams, Fusing Data and Design to Supercharge Innovation-In Products and Processes, McKinsey Global Institute (MGI), 2019 (April).
[14] A.J. Isaksson, I. Harjunkoski, G. Sand, The impact of digitalization on the future of control and operations, Comput. Chem. Eng. 114 (2018) 122–129.
[15] M. Gärtler, V. Khaydarov, B. Klöpper, L. Urbas, The machine learning life cycle in chemical operations – status and open challenges, Chem. Ing. Tech. 12 (2021) 1–19.
[16] H.E.A. Adam, J.K. Kimotho, J.G. Njiri, Multiple faults diagnosis for an industrial robot fuse quality test bench using deep-learning, Results in Engineering 17 (March) (2023), 101007.
[17] Ian Goodfellow, A.C. Yoshua Bengio, The Deep Learning Book 521, MIT Press, 2017, p. 785, 7553.
[18] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436.
[19] F. Lv, C. Wen, Z. Bao, M. Liu, Fault diagnosis based on deep learning, Proc. Am. Control Conf. (2) (2016) 6851–6856.
[20] D. Rolnick, P.L. Donti, L.H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A.S. Ross, N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown, Tackling Climate Change with Machine Learning, 2019. *ArXiv Preprint ArXiv:1906.05433*.
[21] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.
[22] Andrew Ng, Launches A Campaign For Data-Centric AI, 2021. Retreived April 20, 2023 from, https://www.forbes.com/sites/gilpress/2021/06/16/andrew-ng-launches-a-campaign-for-data-centric-ai/?sh=5dea92f374f5.
[23] A. Wu, A Chat with Andrew on MLOps: From Model-Centric to Data-Centric AI, 2021. Retreived April 20, 2023 from, https://www.youtube.com/watch?v=06-AZXmwHjo.
[24] Data Centric AI, Retreived April 20, 2023 from, https://www.datacentricai.cc/, 2021.
[25] Data-Centric AI Competition, Retreived April 20, 2023 from, https://deeplearning-ai.github.io/data-centric-comp/, 2021.
[26] Data centric AI Day, Retreived April 20, 2023 from, https://www.data-centric-ai.com/, 2021.
[27] E. Blasch, T. Pham, C.-Y. Chong, W. Koch, H. Leung, D. Braines, T. Abdelzaher, Machine learning/artificial intelligence for sensor data fusion–opportunities and challenges, IEEE Aero. Electron. Syst. Mag. 36 (7) (2021) 80–93.
[28] J. Gao, P. Li, Z. Chen, J. Zhang, A survey on deep learning for multimodal data fusion, Neural Comput. 32 (5) (2020) 829–864.
[29] J. Li, D. Hong, L. Gao, J. Yao, K. Zheng, B. Zhang, J. Chanussot, Deep learning in multimodal remote sensing data fusion: a comprehensive review, Int. J. Appl. Earth Obs. Geoinf. 112 (2022), 102926.
[30] J. Liu, T. Li, P. Xie, S. Du, F. Teng, X. Yang, Urban big data fusion based on deep learning: an overview, Inf. Fusion 53 (2020) 123–133.
[31] S.R. Stahlschmidt, B. Ulfenborg, J. Synnergren, Multimodal deep learning for biomedical data fusion: a review, Briefings Bioinf. 23 (2) (2022), bbab569.
[32] Z. Li, L. Tian, Q. Jiang, X. Yan, Fault diagnostic method based on deep learning and multimodel feature fusion for complex industrial processes, Ind. Eng. Chem. Res. 59 (40) (2020) 18061–18069.
[33] N. Wang, F. Yang, R. Zhang, F. Gao, Intelligent Fault diagnosis for chemical processes using deep learning multimodel fusion, IEEE Trans. Cybern. 52 (7) (2022) 7121–7135.
[34] X. Xu, Z. Tao, W. Ming, Q. An, M. Chen, Intelligent monitoring and diagnostics using a novel integrated model based on deep learning and multi-sensor feature fusion, Measurement 165 (2020), 108086.
[35] B. Bigdeli, P. Pahlavani, H. Amini, An ensemble deep learning method as data fusion system for remote sensing multisensor classification, Appl. Soft Comput. 110 (2021), 107563.
[36] M.S. Rahman, M. Kaykobad, On Hamiltonian cycles and Hamiltonian paths, Inf. Process. Lett. 94 (1) (2005) 37–41.
[37] C.B. Hurley, R.W. Oldford, Pairwise display of high-dimensional information via eulerian tours and Hamiltonian decompositions, J. Comput. Graph Stat. 19 (4) (2010) 861–886.
[38] M. Elhefnawy, A. Ragab, M.-S. Ouali, Polygon generation and video-to-video translation for time-series prediction, J. Intell. Manuf. (2022) 1–19.
[39] A. Bathelt, N.L. Ricker, M. Jelali, Revision of the Tennessee eastman process model, IFAC-PapersOnLine 48 (8) (2015) 309–314.
[40] Y. LeCun, Y. Bengio, others, Convolutional networks for images, speech, and time series, in: The Handbook of Brain Theory and Neural Networks 3361, 1995, p. 1995, 10.
[41] T. Yu, H. Zhu, Hyper-parameter Optimization: A Review of Algorithms and Applications, 2020. *ArXiv Preprint ArXiv:2003.05689*.
[42] T. Elsken, J.H. Metzen, F. Hutter, Neural architecture search: a survey, J. Mach. Learn. Res. 20 (1) (2019) 1997–2017.
[43] A. Ragab, M. Amazouz, Decision Fusion for Fault Classification in Industrial Processes. *Annual Reliability and Maintainability Symposium (RAMS)*, 2021.
[44] W. Bounoua, A. Bakdi, Fault detection and diagnosis of nonlinear dynamical processes through correlation dimension and fractal analysis based dynamic kernel PCA, Chem. Eng. Sci. 229 (2021), 116099.
[45] K.E. Pilario, M. Shafiee, Y. Cao, L. Lao, S.-H. Yang, A review of kernel methods for feature extraction in nonlinear process monitoring, Processes 8 (1) (2019) 24.
[46] R. Tan, T. Cong, J.R. Ottewill, J. Baranowski, N.F. Thornhill, An on-line framework for monitoring nonlinear processes with multiple operating modes, J. Process Control 89 (2020) 119–130.
[47] R. Tan, J.R. Ottewill, N.F. Thornhill, Nonstationary discrete convolution kernel for multimodal process monitoring, IEEE Transact. Neural Networks Learn. Syst. 31 (9) (2019) 3670–3681.
[48] A. Stief, R. Tan, Y. Cao, J.R. Ottewill, N.F. Thornhill, J. Baranowski, A heterogeneous benchmark dataset for data analytics: multiphase flow facility case study, J. Process Control 79 (2019) 41–55.
[49] R.N. Bracewell, R.N. Bracewell, The Fourier Transform and its Applications vol. 31999, McGraw-Hill, New York, 1986.
[50] L. Aguiar-Conraria, M.J. Soares, The continuous wavelet transform: moving beyond uni- and bivariate analysis, J. Econ. Surv. 28 (2) (2014) 344–375.
[51] D. Granda, W.G. Aguilar, D. Arcos-Aviles, D. Sotomayor, Broken bar diagnosis for squirrel cage induction motors using frequency analysis based on MCSA and continuous wavelet transform, Math. Comput. Appl. 22 (2) (2017) 30.
[52] M. Jalayer, C. Orsenigo, C. Vercellis, Fault detection and diagnosis for rotating machinery: a model based on convolutional LSTM, Fast Fourier and continuous wavelet transforms, Comput. Ind. 125 (2021), 103378.
[53] S.M.K. Zaman, H.U.M. Marma, X. Liang, Broken rotor bar fault diagnosis for induction motors using power spectral density and complex continuous wavelet transform methods, in: 2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE), 2019, pp. 1–4.
[54] L. Eren, T. Ince, S. Kiranyaz, A generic intelligent bearing fault diagnosis system using compact adaptive 1D CNN classifier, J. Sig. Process. Syst. 91 (2) (2019) 179–189.
[55] S. Inazumi, S. Intui, A. Jotisankasa, S. Chaiprakaikeow, K. Kojima, Artificial intelligence system for supporting soil classification, Result. Eng. 8 (2020), 100188.
[56] C. Jing, J. Hou, SVM and PCA based fault classification approaches for complicated industrial process, Neurocomputing 167 (2015) 636–642.
[57] R.D. King, C. Feng, A. Sutherland, Statlog: comparison of classification algorithms on large real-world problems, Appl. Art. Intell. Int. J. 9 (3) (1995) 289–333.
[58] A. Kummer, T. Ruppert, T. Medvegy, J. Abonyi, Machine learning-based software sensors for machine state monitoring - the role of SMOTE-based data augmentation, Result. Eng. 16 (August) (2022).
[59] R. Niyirora, W. Ji, E. Masengesho, J. Munyaneza, F. Niyonyungu, R. Nyirandayisabye, Intelligent Damage Diagnosis in Bridges Using Vibration-Based Monitoring Approaches and Machine Learning: A Systematic Review. In Results in Engineering vol. 16, Elsevier B.V, 2022, 100761.
[60] X. Qi, Z. Yuan, X. Han, Diagnosis of misalignment faults by tacholess order tracking analysis and RBF networks, Neurocomputing 169 (2015) 439–448.
[61] A. Ragab, M. El Koujok, H. Ghezzaz, M. Amazouz, M.-S. Ouali, S. Yacout, Deep understanding in industrial processes by complementing human expertise with interpretable patterns of machine learning, Expert Syst. Appl. 122 (2019) 388–405.

[62] A. Ragab, M.S. Ouali, S. Yacout, H. Osman, Remaining useful life prediction using prognostic methodology based on logical analysis of data and Kaplan–Meier estimation, J. Intell. Manuf. 27 (5) (2016) 943–958.
[63] J. Tian, C. Morillo, M.H. Azarian, M. Pecht, Motor bearing fault detection using spectral kurtosis-based feature extraction Coupled with K-Nearest Neighbor distance analysis, IEEE Trans. Ind. Electron. 63 (3) (2016) 1793–1803.
[64] Q. Zhang, J. Gao, H. Dong, Y. Mao, WPD and DE/BBO-RBFNN for solution of rolling bearing fault diagnosis, Neurocomputing 312 (2018) 27–33.
[65] W. Zhou, X. Li, J. Yi, H. He, A novel UKF-RBF method based on adaptive noise factor for fault diagnosis in pumping unit, IEEE Trans. Ind. Inf. 15 (3) (2019) 1415–1424.
[66] A. Ragab, M. El Koujok, H. Ghezzaz, M. Amazouz, in: J. Ren, W. Shen, Y. Man, L. B, T. A, A. I, P.S.E. Dong (Eds.), Chapter 10 - Fault Diagnosis in Industrial Processes Based on Predictive and Descriptive Machine Learning Methods, 254, Elsevier, 2021, p. 207.
[67] K.K. Dobbin, R.M. Simon, Optimally Splitting Cases for Training and Testing High Dimensional Classifiers, 2011.
[68] Q.H. Nguyen, H. Ly, L.S. Ho, N. Al-ansari, H. Van Le, V.Q. Tran, I. Prakash, B. T. Pham, Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil, 2021.
[69] B.T. Pham, I. Prakash, A. Jaafari, D.T. Bui, Spatial prediction of rainfall-induced landslides using Aggregating one-dependence estimators classifier, J. Indian Soc. Remote Sens. 46 (9) (2018) 1457–1470.
[70] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, Tensorflow: a system for large-scale machine learning, in: 12th Symposium on Operating Systems Design and Implementation, 2016, pp. 265–283.
[71] PySpark, Retreived April 20, 2023 from, https://spark.apache.org/docs/latest/api/python/index.html, 2023.

Dr. Mohamed Elhefnawy received a BSc in Communication & Information Engineering at the University of Science and Technology, Zewail City in Egypt in 2018. He received a Ph.D. degree in Industrial Engineering from Polytechnique Montréal, the University of Montréal, Canada in 2022. His main research interests include Artificial Intelligence (AI), Machine Learning, Deep Learning (DL), IoT, Operations Research, Decision Sciences & Statistics. Dr. Elhefnawy is interested in applying AI in different applications to maximally exploit the available data for a better decision-making process. During his Ph.D., he worked as a research assistant for data integration & fusion at the CanmetENERGY research center of Natural Resources Canada. At CanmetENERGY, he has worked on strategic projects funded by the Government of Canada for Forest Value Chain Optimization (FVCO). Dr. Elhefnawy has developed AI-based methodologies to optimize the operation of industrial processes. The aim is to improve the efficiency of energy-intensive processes that include complex industrial equipment such as recovery boilers, concentrators, heat exchangers, etc. After getting his Ph.D. degree in February 2022, Dr. Elhefnawy joined the Integrated Remote Sensing Studio (IRSS) lab at the University of British Columbia (UBC) as a postdoctoral fellow to apply advanced DL algorithms and tools to exploit next-generation remote sensing data in forestry.

Dr. Mohamed-Salah Ouali is a Professor of Industrial in the Department of Mathematics and Industrial Engineering at Polytechnique Montréal, Québec, Canada, since 2000. He obtained his Doctorate degree from the Institut National Polytechnique de Grenoble, France, in 1996, and worked as an adjunct professor at Moncton University, New Brunswick, Canada, from 1998 to 2000. He is a member of the Interuniversity Research Center on Enterprise Networks, Logistics and Transport (CIRRELT) and the Institute for Data Valorization (IVADO). His research interests focus on reliability and maintenance strategies modeling, multiple failure modes diagnosis, causality analysis, system supervision/control, component aging, multivariate time series analysis, residual life prognosis, condition-based maintenance and its tools, statistical and Bayesian learning, maintenance digital twins, and quantitative risk analysis.

Dr. Ahmed Ragab is an AI scientist at CanmetENERGY, an energy innovation center of Natural Resources Canada. He is also an Adjunct Professor at the Department of Mathematics and Industrial Engineering, Polytechnique Montréal. He is on leave from the Faculty of Electronic Engineering, Menoufia University, Egypt. He received a Ph.D. degree in Industrial Engineering from Polytechnique Montréal, the University of Montréal, Canada in 2014. His research interests include Artificial Intelligence (AI), Machine Learning, Pattern Recognition, Image Processing, Data & Decisions Fusion, Causality Analysis, Reliability Modeling, Operations Research, Discrete Event Systems, and Process Mining. His main thematic activities focus on the practical challenges of Big Data and AI in a number of applications including Abnormal Events Diagnosis & Prognosis, Predictive Maintenance, Supervisory Control, Real-time Optimization, Production Scheduling, and Systems Design. Dr. Ragab has a bunch of experience in developing advanced AI algorithms and tools in the manufacturing industry, aiming at reducing energy consumption, GHG emissions, and operational and maintenance costs while improving operations' performance. He also teaches industrial engineering courses at graduate levels and co-supervises MSc and Ph.D. students.

Dr. Mouloud Amazouz holds a Ph.D. degree in Mechanical Engineering from Polytechnique Montréal. He is a Senior Project Manager at CanmetENERGY, an energy innovation center of Natural Resources Canada. He has more than 30 years of experience in leading, performing and managing research & development and technology demonstration and deployment in the areas of process design and operation optimization through data analytics and artificial intelligence. He co-authored three US patents and more than 50 journal and conference papers. He has strongly contributed to the development, testing, transfer, and commercialization of several energy technologies (hardware and software) in the industry. He also taught mechanical engineering courses at graduate and undergraduate levels and co-supervised MSC and Ph.D. students. He has strong abilities to prepare and negotiate agreements for technology licensing, grants & contributions, services, and R&D.