Intelligent Systems with Applications 16 (2022) 200112

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

Review

Multimodal deep learning for predicting the choice of cut parameters in the milling process

Cheick Abdoul Kadir A Kounta [a,b,*], Bernard Kamsu-Foguem [a], Farid Noureddine [a], Fana Tangara [b]

[a] Laboratoire Génie de Production, École Nationale d'Ingénieurs de Tarbes, 47 Avenue Azereix, B.P. 1629, F-65016 Tarbes Cedex, France
[b] Faculté des Sciences et Techniques, Université des Sciences, des Techniques et des Technologies de Bamako (USTTB), B.P. E : 423, Bamako, Mali

ARTICLE INFO

Keywords:
Deep learning
Multimodal data processing
Heterogeneous data fusion
Unstructured data
Manufacturing processes
Roughness

ABSTRACT

In this paper, we use multimodal deep learning to predict the choice of optimal cutting parameters (cutting speed, depth of cut, and feed rate per tooth) and the appropriate cutting tool for reproducing an existent piece of the same surface state, considering the footprints left by the cutting tool. We use the image of the aluminum plate's surface states considering the tool's footprints, the cutting parameters, and the roughness average (Ra) obtained with a roughness meter to drive our model. We built a late multimodal fusion model with two networks, a convolutional neural network (CNN) and a recurrent neural network with long short-term memory layers (LSTM). The first network consists of the first branch with a convolutional network that receives the input images. In the second network, modeling is performed by the LSTM network to receive the digital input data. This provides a framework to integrate information from two modalities to ensure surface quality in machining processes. This approach aims to assist in selecting the appropriate cutting tool and cutting parameters to automatically reproduce a machined piece using the image and roughness of an already existing piece. It is observed that the performance of the multimodal model is better than that of the unimodal model on image data. The accuracy continues to improve on both sets (training and validation), and the multimodal model finally reaches good accuracy results. Contrary to the unimodal model, which fails to generalize the training on a validation dataset. The results estimated by the multimodal fusion model are encouraging when applied to the milling activity in industrial production processes.

1. Introduction

The advancement of sensor technology in the industry has led to the automation of machining process tasks, which has resulted in the increasing use of sensors for mechanized production lines (Broo et al., 2021). The automation of these tasks has spread beyond the automotive and manufacturing industries. Industry 4.0 is based on the Internet of things, recognition solutions, cybersecurity, integration of horizontal and vertical systems, cloud computing, 3D printing, additive manufacturing, big data, business analysis, augmented reality, and machine learning (de la Peña Zarzuelo et al., 2019). The development of deep learning combined with image capture techniques provides powerful tools for recognizing the features and information contained in an image. This leads the way for the automation of image content analysis, possibly extracting relevant functionality. The learning systems used are based on artificial neurons with architectures whose entities are linked together to form many layers of varying depth. In the first layers, these architectures allow the hierarchical processing of data on several scales. In addition, the deeper the layers, the more invariable the network's responses.

These deep neural networks can therefore play an essential role in the analysis of multimodal data, and in particular for a better understanding of the characteristics describing the surface quality levels of objects studied. Multimodal fusion has generated great interest in multimodal data processing and general concern among researchers (Bayoudh et al., 2021). Multimodal fusion includes early feature-based fusion, late decision-based fusion, and hybrid fusion combining the two previous modes (Baltrušaitis et al., 2017). To solve the problem of the discrepancy between the information from the original data of each modality before learning, feature-level multimodal fusion is used to fuse the extracted features from each modality. regarding the decision-level fusion method, it integrates knowledge generated from multimodal

* Corresponding author at: Laboratoire Génie de Production, École Nationale d'Ingénieurs de Tarbes, 47 Avenue Azereix, B.P. 1629, F-65016 Tarbes Cedex, France.
E-mail address: ckounta@enit.fr (C.A.K.A. Kounta).

https://doi.org/10.1016/j.iswa.2022.200112
Received 8 April 2022; Received in revised form 19 July 2022; Accepted 10 August 2022
Available online 18 August 2022

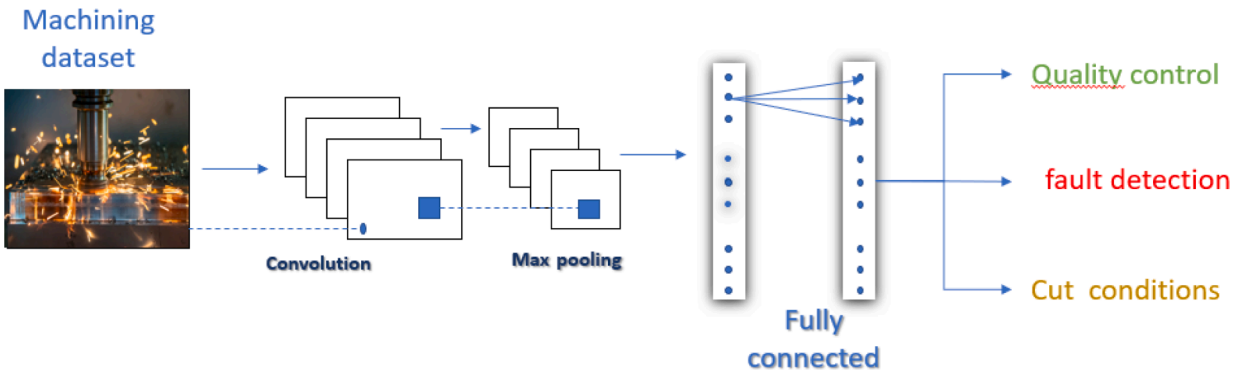

Fig. 1. Basic CNN model.

information into different basic models, and then a balanced fusion decision is made for the results of these models. By combining feature-level fusion and decision-level fusion, hybrid multimodal fusion has a more flexible structure. The machining of pieces is a constant need for industrial production systems. The reproduction of certain original pieces or the implementation of new pieces is a manufacturing technique that must take into account at the same time the surface roughness, the cutting conditions, the cutting tool, the dimensions, and the vibration signals (Hu et al., 2019). In this context, the three main components of CNC machining are a machine tool, a workpiece, and a cutting tool. Specifically, milling is the removal of material from a workpiece by the cutting tool, which is accompanied by a series of changes in geometric shape and physical parameters, such as cutting resistance, cutting temperature, tool wear, etc. When planning the machining process, the cutting tool and the cutting parameters are decisive elements. Proper selection of cutting tools and cutting parameters can significantly reduce energy consumption and production time for the machining process (Chen et al., 2019; Okokpujie et al., 2018).

Reducing the energy consumption of machine tools for turning operations is important in promoting sustainable manufacturing. Studies have shown that selecting optimal cutting parameters is an effective approach to reducing cutting energy consumption in machining (Chen et al., 2021). To illustrate this, particular interest is given to milling, which is a manufacturing process in which the removal of material in the form of chips results from the combination of two movements: the rotation of the cutting tool, on the one hand, and the advancement of the workpiece on the other. In milling, surface quality is the most important factor in the assessment of the manufactured components.

This approach aims to help select the appropriate cutting tool and cutting parameters to automatically reproduce a machined piece from its image and its roughness. First, we start with the acquisition of delicate data, and the constitution of databases from heterogeneous elements. In this database, links have been established between the modalities to facilitate the following steps of the proposed methodology of fusion in the considered industrial context. The originality of our proposal lies on two levels: i) the alignment of textual data with annotated images for a better multimodal data acquisition; ii) the combination of several modalities to improve the prediction of crucial parameters (cutting speed, depth of cut, and feed rate per tooth) to ensure a good surface quality.

After this introduction, Section 2 presents some state of the art in architectures and optimization functions of deep learning and multimodal applications. Section 3 shows the proposed methodology for multimodal fusion learning. A case study is detailed in Section 4. Section 0 consists of analyzing and discussing the results. Finally, Section 6 delivers a conclusion.

2. State of the art

2.1. Deep learning: architectures and optimization

Based on machine learning, deep learning enables layered computational networks to learn and represent data with multiple levels of abstraction. Deep learning, therefore, uses several successive transformations, characteristics, and representations, mimicking the way the brain learns and understands multimodal information, which automatically captures the complex structures of large-scale data (Litjens et al., 2017). This is an emerging approach that has been widely applied in traditional fields of artificial intelligence, such as semantic analysis (Bouwmans et al., 2019), transfer learning (Lu et al., 2015; Palade et al., 2021), natural language processing (Hochreiter and Schmidhuber, 1997), computer vision (Wang et al., 2018).

Deep learning has aroused the interest of researchers for the following reasons: firstly, the increase in computing power with the advent of the graphics processor, access to affordable hardware, the quality of computing platforms, and the increase in network connection speed, which considerably reduces the execution time of algorithms. Secondly, its ability to cope with the increasing amount of data. Thirdly, the extraction of features in an unsupervised way, i.e., without human expertise (Zhao et al., 2019). Deep learning is developing across a large family of methods, encompassing neural networks, hierarchical probabilistic models, and a variety of unsupervised and supervised feature learning algorithms.

There are several deep neural network architectures, such as convolutional networks (CNNs), which are notable in computer vision applications. In these networks, the layers are formed robustly. Deep learning models are widely used in troubleshooting (Duan et al., 2018; Chen et al., 2019; Siyu et al., 2020), prediction of residual life (Yan et al., 2018; Qin et al., 2020; Luo and Zhang, 2022), and monitoring the condition of mechanical equipment (Xiang et al., 2022).

A CNN consists of three main layers, namely convolutional layers, pooling layers, and fully connected layers, with each layer playing a different role (Indolia et al., 2018). The former layers extract generic features, the latter layers extract specific features, and the goals pursued may be for classification (discrete output values) or regression (continuous output values). In a CNN, the extracted features are learned by hierarchical sampling layers and stacked by reducing the number of parameters. Hierarchical model learning has high computational complexity (Wen et al., 2015). A basic model showing the process of running a CNN is shown in Fig. 1. This network receives as input a series of images whose features are extracted by convolution layers. Subsampling is performed between the convolution layers by the max-pooling layer to reduce the image size for the next convolution layer by keeping only the most significant features. By stacking multiple convolutions and max-pooling layers, a fully connected layer is applied to decide at a higher level in the network. The fully connected layer makes connections to all outputs of the previous layers. Then the

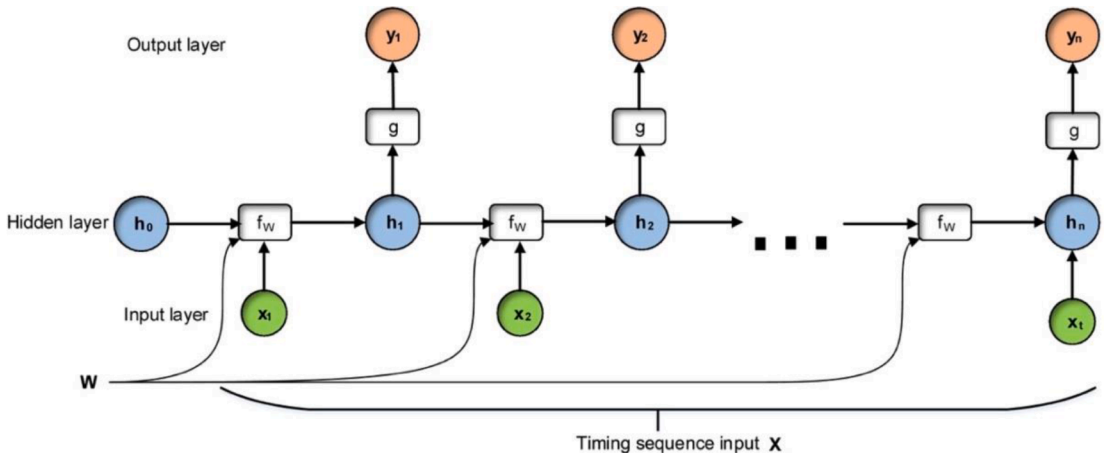

Fig. 2. Architecture of the recurrent neural network (Wang et al., 2018).

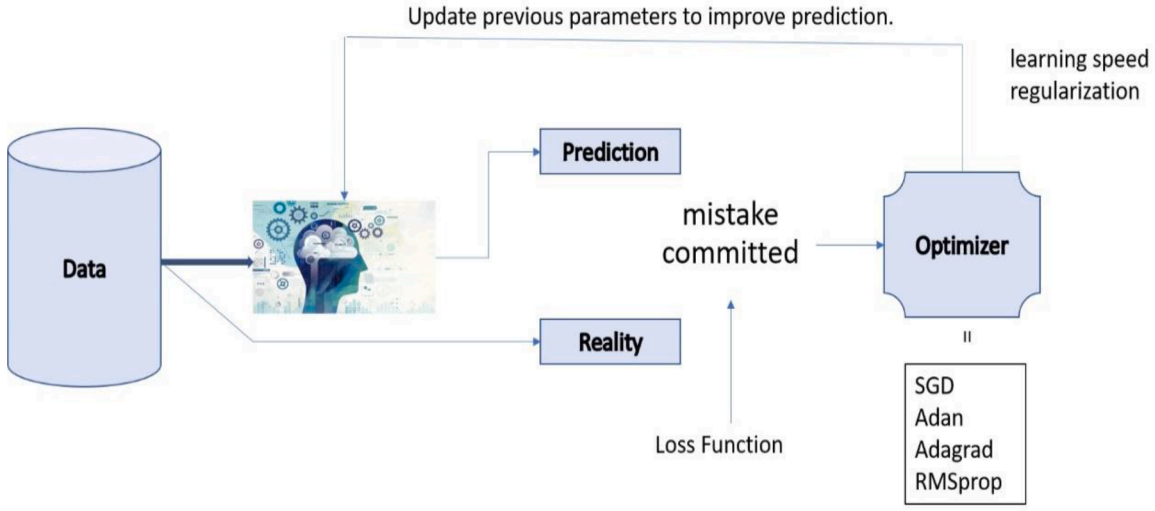

Fig. 3. How the optimizer works on an artificial intelligence model.

activation functions are computed in networks by multiplying the matrices. Finally, the difference between the prediction and the reality is calculated with the loss function. The loss functions are used to make the prediction. The number of layers in these models depends on the complexity of the input data.

Recurrent networks are neural networks in which information is distributed back and forth, from the deep layer of the first layer. Through recurring connections, these networks can keep information in memory and take into account several past states at any given time. For this reason, RNNs are more suitable for the processing of temporal sequences such as learning and signal generation (Emmert-Streib et al., 2020). RNN can only store in the near past and start to forget after about 50 iterations. This two-way transfer of information makes it much more difficult to learn, and it is only recently that effective methods have been developed, such as LSTM (Long Short-Term Memory). These large long short-term memory networks have revolutionized speech recognition (Graves et al., 2013) or the understanding and generating of Natural Language Processing text (Wen et al., 2015).

Recurrent networks are neural networks in which information is distributed in both directions, from the deep layer to the first layer. Thanks to recurrent connections, these networks can keep information in memory and take into account several past states at a given time. For this reason, RNNs are best suited for processing temporal sequences such as learning and signal generation. RNNs apply a rule to update the hidden state calculation at every step. If we take the example in Fig. 2, the sequential input is considered a vector. We can therefore calculate the current hidden state in two phases through an activation function. The first party input layer is calculated with the input while the second party is obtained from the hidden layer in the previous step. The prediction can be computed with the current hidden layer via a softmax function.

To improve the model, we will compare the predictions of the model with the real values: this is the loss function of the model. An algorithm called an optimizer then plays the role of correcting the model algorithm (for example the weights of a perceptron) so that the next time, it can predict a value closer to reality.

The role of the optimizer is critical, it will define how the AI system evolves (learns) to adapt itself to the training data. Fig. 3 shows how an optimizer works.

2.2. *Multimodality in industrial applications*

There are several methodologies capable of integrating information from multimodal sensors to improve monitoring, diagnostics, and prediction performance. Multimodal data fusion is a very dynamic field of research applied in various industrial sectors and their interdisciplinary areas such as automation, manufacturing, and robotics (Bokade et al., 2021). The main objective is to process information from several

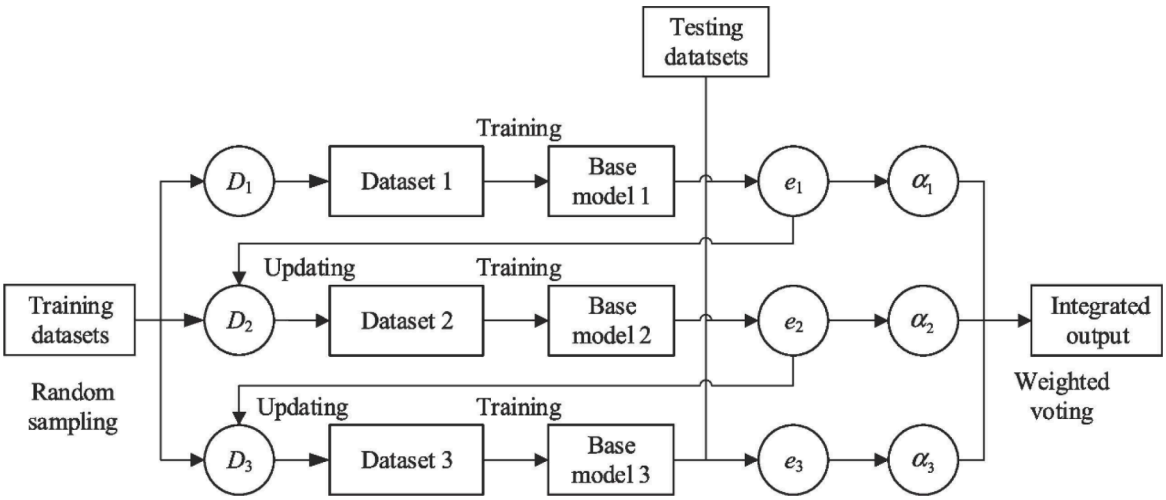

Fig. 4. : the model of multimodal decision fusion based on whole learning (Che et al., 2020).

heterogeneous sources to estimate an almost precise view of the structural, functional, and behavioral states of a machine. The observations of these states provide elements of analyses and reflections to specialists and professionals in the industry. It can offer a range of reasoning tools with sometimes their digital interactive service in the recognition of the scene (Gupta et al., 2021; Alkhalaf, 2021) and human-robot interaction (Cuayáhuitl, 2020; Liu et al., 2018). Indeed, in some complex environments, it is necessary to use multiple modalities that provide additional information about the same situation and offer a great opportunity for the resolution of problems. The mechanism of a multimodal procedure has also been shown to be useful for the detection and diagnosis of defects (Ma et al., 2018), failure prognostics (Yang et al., 2021), Remaining Useful Life (RUL) estimation (Al-Dulaimi et al., 2019) of systems in industrial organizations. The approach based on Remaining Useful Life (RUL) data can be useful in estimating cutting tools during the milling process (Kumar et al., 2022). One can imagine a use of the RUL in a complementary way to our approach to tighten or relax the tolerance constraints of the cutting conditions. In the case of a large RUL reflecting a long-life expectancy of the tool, the requirements can be high, while in the case of a small RUL reflecting a low life expectancy of the tool and aging of the tool requirements can be mitigated.

In the context of a diagnosis of defects, the authors suggest a deep coupling autoencoder (DCAE) model, which captures the multimodal sensory signals belonging to a measurable space, such as sound and vibration data, and incorporates feature extraction of multimodal data effortlessly into data fusion for the diagnosis of fault modes (Ma et al., 2018). In the context of failure prognostics, the authors have proposed a method based on the construction of a multi-branch DNN containing three branches: (i) The imaging branch, which uses a CNN model to extract characteristics from an image; (ii) the text branch, extracting characteristics from inspection records using a CNN model and (iii) the numerical data branch, having a vector of numerical information employs feedforward neural networks (Yang et al., 2021). Regarding the Remaining Useful Life (RUL) estimation, the authors have suggested a Hybrid Deep Neural Network Model (HDNN) structure that comprises two parallel models (one LSTM and one CNN) preceding a fully connected multilayer neural network fusing the output of each model to form the RUL pursued. The LSTM is employed to extract temporal characteristics while concurrently the CNN is used to extract spatial characteristics, which are then combined for giving the global prediction of the RUL. To solve the problems of conventional roller bearing longevity prediction methods, which focus on prediction accuracy by ignoring the cost and time, the authors propose a new prediction approach by fusing a one-dimensional convolutional neural network and a simple recurrent neural network. To extract features from the signal, they use the functionality of the one-dimensional convolutional neural network. The maximum global pooling layer is used to replace the fully connected layer. In the prediction part, a parallel input network was established to build the serial operation fusion model of a traditional recurrent neural network (RNN) (AlZubi et al., 2021). Other authors propose a multimodal characteristic extraction plan to get hierarchical representations from the original vibration signal and the current envelope. First, they form two fully connected (FC) layers to fast reduce the input dimension and learn the primary characteristics. Then, two convolutional layers are piled up to learn a more compact and better representation of the characteristics of each model. The extracted features are then remodeled into serval vectors, and every vector will implicitly encode a fault characteristic. Eventually, the dynamic routing algorithm is implemented as the integration of the fusion module and the classifier. During the training phase, the bias-injected vibration signal and the converted current signal are fed into the network at different inputs. Next, the dynamic routing procedure relates the various sources to the different fault models and then performs the predictions (Fu et al., 2020).

Multimodal machine learning has some challenges such as representation, translation, alignment (Yu et al., 2022), fusion (Poria et al., 2016), and co-learning (Rahate et al., 2022). These challenges appear different despite the existence of a crossover phenomenon. They are often applied jointly to perform multiple operations using multimodal deep learning models. For example, to perform a multimodal fusion, it is necessary to apply a representation to acquire additional and complementary information. Some authors have mentioned the use of multimodality in the industrial context. A deep learning method based on the fusion of multimodal functionalities for the online diagnosis of rotating machines has been presented by (Zhou et al., 2018). A simulation was carried out and a practical case study was conducted to validate the effectiveness of the method. The proposed method was compared to the method based on the convolutional neural network without feature fusion. An increase of approximately 5% in the performance of the fusion method compared to the non-fusion method was observed. The interest in multimodal learning stems from the three main advantages it can provide.

First, having access to multiple modalities that observe the same phenomenon can lead to more robust predictions. Second, having access to multiple modalities might allow us to capture additional information which is not noticeable independently in the individual modalities. Third, a multimodal system can still function when one of the modalities is missing, for example, it is possible to recognize the characteristics of a phenomenon with the visual signal in the absence of an audio signal. We can see in Fig. 4 a multimodal fusion model used for the diagnosis of

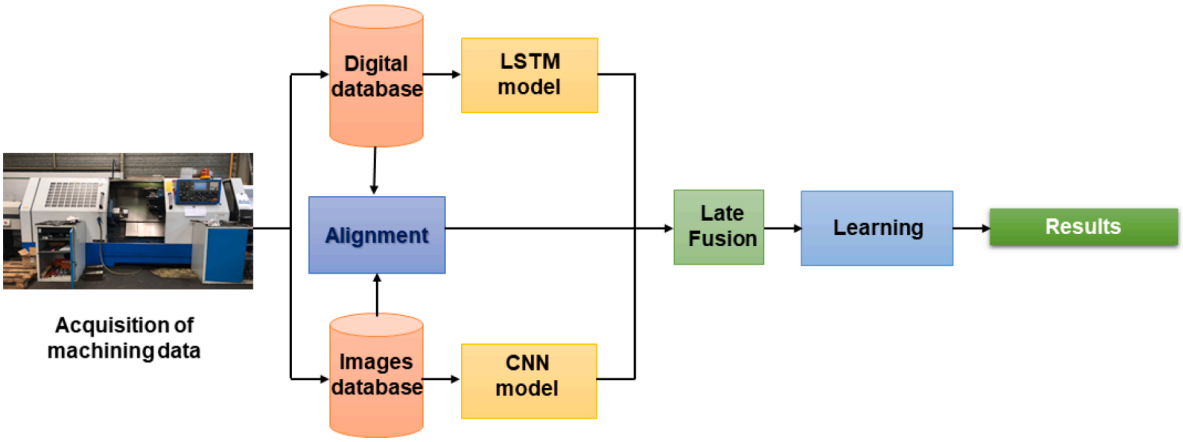

Fig. 5. Diagram of the proposed methodology.

bearing failure, which is composed of four parts: multimodal data acquisition, multimodal fusion at the functionality level, training and optimization of deep learning models, and multimodal merger at the decision level.

The grayscale image samples, and time series samples are fed into CNN and Deep Belief Network (DBN), respectively, to form several hidden layers and optimize the model parameters, to obtain basic classifiers for bearing fault diagnosis. The complete fault diagnosis results are obtained by multimodal decision fusion, which is performed by ensemble learning of several different deep learning models. The CNN, which is commonly applied in image processing, is used to process grayscale images, and the DBN is used to form time-series samples. At the decision level, the combination of different deep learning models aims to achieve comprehensive fault diagnosis results. To predict the cutting force of face milling, one modeling approach is to provide information and knowledge about the relationship between cutting forces and cutting conditions such as cutting speed, depth of cut, and cutting speed advance (Charalampous, 2021). The experiments were performed in a three-axis CNC milling center. The milling trials were performed at random by collecting the dataset to create the predictive cutting force model. The application's evaluation of this model was to effectively calculate the cutting forces and suggest cutting conditions that would reduce the stress fields developed during milling. This review of the literature provides some inspiring observations. Some theoretical work on artificial intelligence aims to design and implement new methods. The objective is to improve certain methods, to make them more robust, more general, faster, or to reduce their complexity. Another work that seeks to apply AI or big data algorithms to meet application needs or societal challenges is observed in the literature that most actions dealing with multimodality are oriented towards medical and societal applications, but there is very little work related to the industrial framework.

We also note that the potential of multimodal machine learning is still in its infancy in the industrial world, but its relevance is not in doubt. This is particularly underlined by recent work using a multi-branch neural network for the prognosis of failures based on multimodal data (Yang et al., 2021). In terms of fusion strategies, the early fusion strategy raises questions about the impact of this concatenation of heterogeneous data on the performance obtained. Therefore, we will opt for a late fusion strategy that performs a separate learning model for each modality before fusion these models to obtain the final classification. In addition, the interaction with experts in the industrial field, allowed us to introduce a priori information for the determination of the usual machining conditions. Incorporating this information is likely to have a positive impact on the training model process.

3. Proposed methodology for multimodal fusion learning

3.1. Steps of the methodology

In this section, we explain the proposed work methodology. It includes six steps, namely data collection and preparation, choice of processing models, model fusion, training of the fused models, and interpretation of the result.

- Data collection and preparation: the data is collected from a milling process on an aluminum plate and put into an appropriate format required for processing. These experiments have resulted in two types of image and digital datasets.
- Alignment: it is a process of matching elements from two or more modalities. The relationship between each image from the experiment and their cutting conditions (digital database) is determined in this phase.
- Choice of treatment models: this step consists in creating a model adapted to the treatment of each type of data. In this case, a CNN is proposed to process images and an RNN with LSTM layers for digital data.
- Fusion of models: it allows us to take advantage of several modalities to solve the same problem. A late fusion model is proposed, i.e., fusion at the model level to perform training.
- Model training: This is the specific aspect of deep learning with the constructed multimodal neural architecture. This step consists in determining the loss function, the optimizer, theme, tricks, and the training time of the fusion model.
- Interpretation of the result: it consists of providing information on the result obtained concerning the data used and the learning process: the training and validation curves.
- The results are interpreted using the most popular performance indicators: classification accuracy and loss rate. The accuracy allows for the simple and efficient transmission of information in numbers. The loss rate is an essential complement to the accuracy. The loss function is based on the calculation of the cross-entropy. It quantifies the difference between two probabilistic distributions. A negligible loss rate represents a better model performance. The cross-entropy function of the loss is shown in functions 1 and 2.

$$L = \frac{1}{n}\sum_{i-1}^{n} y_i log(\widehat{y}_i) + (1 - y_i) log(1 - \widehat{y}_i) \#(1)$$

$$L = -\sum_{i=1}^{n} y_i log(\widehat{y}_i) \#(2)$$

Fig. 6. Process of fusion of two networks LSTM and CNN.

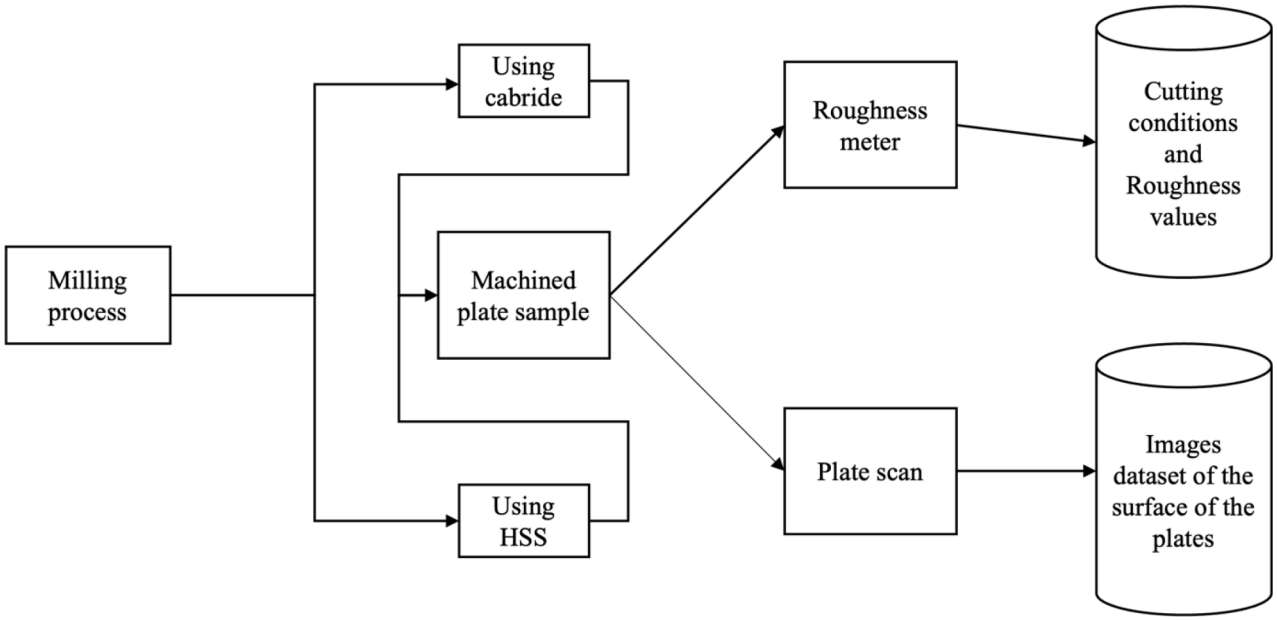

Fig. 7. The data collection process on the milling machine.

The output layer of the model is defined with n nodes (one node for each class) and the closer the value of the loss function is to 0, the more reliable it is.

The accuracy is the ratio between the number of correctly predicted classes and the total number of examples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \#(3)$$

$$Accuracy(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} 1(\hat{y}_i = y_i) \#(4)$$

$\hat{y}_i$ is the predicted value of the ith sample, y_i represents the actual value and n the number of samples. Where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives.

These steps of this approach are schematically shown in Fig. 5.

3.2. Learning models of the methodology

3.2.1. CNN model

A normalization process was performed on the image data in the learning phase of the network. Since the raw images have a variable dimension and each pixel of the RGB channels is in the range [0, 255]. This normalization will scale the pixel values to the same level. In the deep learning model, the output of each layer is the input of the next. The normalization at the input will make the learning easier and to get a better learning performance it is done at the input of each layer.

The model consists of five CNN layers, which are used to extract features from the image. At the output of each convolution layer, a normalization process is performed using the batch_normalization layer. The stacked CNN layers are analyzed by a maximum pooling layer, which reduces the data dimensions by combining the outputs of the neural clusters in one layer into a single neuron in the next layer. The CNN layers consist of 3 filters of size (3 × 3). The maximum filter size of the clustering layers is (2 × 2). We add two more layers, a dropout layer to turn off the random parts of the neurons so that the network adapts to the lack of information and a flattened layer is added to gather all the images (matrices) we have into one (long) vector. The CNN model receives as input a series of images of dimension 250*166 by a convolution layer followed by a normalization operation. Then the rest of the convolution layers are followed by max-pooling layers to compress their outputs and retain only the essential information. All the features are kept in a flattened layer prepared for the fusion operation. Fig. 6 shows the detailed architecture of the CNN model of the methodology.

3.2.2. LSTM model

The LSTM model consists of an input layer, and two stacked LSTM layers for feature extraction from digital data. Each of these layers includes a structure of 64 neurons, and a third flattened layer (see Fig. 7). In this part, normalization operation on data has been performed before introduction into the network.

3.2.3. Fusion process

The fusion of the CNN-LSTM models has attracted the attention of researchers in different fields (Kim and Kim, 2019). (Mou et al., 2021) propose a CNN-LSTM fusion model to detect the stress level of a driver. The results of this method outperform the state-of-the-art models with an average accuracy of 95.5%. On the other hand, (Hatami et al., 2022) use LSTM-CNN fusion in the medical field to predict the clinical outcome of stroke patients. To improve efficient extraction and avoid information loss on EEG signals (Li et al., 2022) extracts the spatial features with CNN and temporal features with LSTM and then fuses their outputs. This technique has improved the accuracy of EEG. The LSTM-CNN fusion has proven to be effective in the diagnosis of gearbox failures. This fusion made it possible to determine the type of fault, its location, and its direction (Shi et al., 2022). Based on this literature, a multimodal LSTM-CNN fusion model is designed to process the data.

Among the existing fusion methods, we use fusion at the level of the classifier model called late fusion, as we do feature extraction for both models in parallel and then fuse the outputs of these models forming an end-to-end network with an output vector whose components are the cutting tool types and cutting condition parameters. The first part receives the output of the LSTM model for the numerical data with an output of 192 feature neurons and the second party receives the output of the CNN network for the images with an output of 162688. The outputs of the LSTM model (left side of Fig. 6) and the CNN model (right side of Fig. 6) are merged with the concatenate function and then apply an activation function (relu) on the merge. A fully connected layer is added with 128 units to form the network outputs. Each parameter has its own output layer to perform an evaluation of each output. The prediction of the parameters that have a numerical value is performed by a regression technique using the mean square error (MSE) and the mean

Table 1
Characteristics of the VF 1 machine.

Description	Metric
Maximum spindle speed	8100 rpm
Maximum spindle torque	122.0 Nm at 2000 rpm
Maximum milling	16.5 m/min
Fast traverse speeds X-Y-Z	25.4 m/min
Maximum force X-Y-Z	11343 N
Tool change capability	30
Maximum tool diameter	89 mm
Internal pallet	257 cm x 251 cm x 257 cm
External pallet	249 cm × 232 cm × 254 cm
Spindle speed	8100 rpm
Spindle power	22.4 kW

absolute error (MAE) loss function as metrics. Only the cutting tool is predicted with the softmax activation function (HSS and carbide), the categorical cross-entropy loss function, and the accuracy as metrics. This fusion can be visualized in Fig. 6.

4. Case study

Specialized tooling machines, such as CNC milling machines, have become essential components of machining centers for their ability to machine complex shapes without disassembling the part. The milling process for precision machining involves cutting metal to design precision mechanical parts. The milling process is performed by a CNC assistant and consists of four steps. The first step is to program the computer from the interface using an automatic CAM (computer-aided manufacturing) or CAD (computer-aided design) programming software. The programming process is done by entering a series of numbers and letters after performing the appropriate calculations. Once programmed, the material to be machined and the required cutting tool are placed according to the type of cut to be made the next step is to start the milling process for precision machining. Using deep learning techniques, this system will be able to predict the cutting conditions to reproduce an existing piece by having the images and the arithmetic mean roughness of the piece. Since there are millions of tool combinations, it is impossible to make such predictions by the eye. The approach developed can be used as a decision-support system to determine the appropriate cutting conditions for the reproduction of an existing piece.

4.1. Milling process

The milling process is a machining operation using rotating tools for material on a given piece. In this study, milling was performed on a CNC milling machine (Haas VF1type), data is collected by experimenting on an AU4G aluminum plate (2017A) using different cutting conditions to achieve several surface finishes. Table 1 gives us an overview of the characteristics of the milling machine.

Two cutting tools were used on the plates (tungsten carbide and HSS) to perform the milling (as shown in Fig. 7). Several cutting speeds (120, 140, 160, 180, and 200) were used with a variation of the feed rate per tooth (0.01, 0.02, 0.03, 0.04, and 0.05) for each cutting speed, and the feed rate that remained constant throughout the operation. Each value of the cutting speed was combined with all values of the feed rate per tooth using the cutting tools. These operations produced several surface conditions to serve as the database for this study.

4.2. Data collection and preparation

The samples of aluminum plates obtained during the experimentation made it possible to obtain two types of data, the images of the machined surfaces and the values of the arithmetical mean roughness of these surfaces. Each combination of cutting conditions was used to obtain a surface condition and the arithmetic mean values of the surface roughness are calculated by a roughness meter. This calculation allowed the construction of the numerical database with the values of the cutting parameters (roughness, feed speed, cutting speed, and feed speed per tooth) used. Table 2 and

Table 2
. Variations of input settings as a function of roughness with digitally controlled processing with the HSS tool.

VcFz	120 (m/min)	140(m/min)	160 (m/min)	180 (m/min)	200 (m/min)
0.01 (mm/tooth)	1.136µm	1.16µm	1.052233µm	1.065µm	0.87µm
0.02 (mm/tooth)	1.208µm	1.27µm	1.170333µm	1.121µm	0.957µm
0.03 (mm/tooth)	1.548µm	1.33µm	1.487µm	1.345µm	1.25µm
0.04 (mm/tooth)	1.668µm	1.593µm	1.522µm	1.495µm	1.487µm
0.05 (mm/tooth)	1.76µm	1.614µm	1.61µm	1.5032µm	1.505µm

Table 3
. Variations of input parameters as a function of roughness with digitally controlled milling with the metal carbide tool.

VcFz	120 (m/min)	140(m/min)	160 (m/min)	180 (m/min)	200 (m/min)
0.01 (mm/tooth)	1.142µm	1.146µm	0.981µm	0.956µm	0.8166µm
0.02 (mm/tooth)	1.145µm	1.188µm	1.019µm	0.982µm	0.8161µm
0.03 (mm/tooth)	1.271µm	1.217µm	1.128µm	1.022µm	0.9261µm
0.04 (mm/tooth)	1.312µm	1.482µm	1.257µm	1.254µm	1.0617µm
0.05 (mm/tooth)	1.595µm	1.515µm	1.46µm	1.356µm	1.152µm

VcFz	120 (m/min)	140(m/min)	160 (m/min)	180 (m/min)	200 (m/min)
0.01 (mm/tooth)	1.136 μm	1.16 μm	1.052233 μm	1.065 μm	0.87 μm
0.02 (mm/tooth)	1.208 μm	1.27 μm	1.170333 μm	1.121 μm	0.957 μm
0.03 (mm/tooth)	1.548 μm	1.33 μm	1.487 μm	1.345 μm	1.25 μm
0.04 (mm/tooth)	1.668 μm	1.593 μm	1.522 μm	1.495 μm	1.487 μm
0.05 (mm/tooth)	1.76 μm	1.614 μm	1.61 μm	1.5032 μm	1.505 μm

Table 3 show the possible combinations of cutting conditions with the two cutting tools with the cutting speed (Vc in m/min), the feed speed per tooth (Fz in mm/teeth), and the arithmetic mean roughness. On the line Fz/Vc we find the values of the cutting speed, the column represents the feed rate per tooth and the center represents the values of the arithmetic roughness. For example, when Fz = 0.01 and Vc = 120, then the value of Ra is 1.136. Since the original data are at very different scales, they have been normalized. All the values are reduced between

Table 4
The representation of data from both sources.

Image	Ra (μm)	Fz (mm/ tooth)	Vc (m/min)	Tool
	1.136	0.01	120	HSS
	1.208	0.02	120	HSS
	1.548	0.03	120	HSS
	1.065	0.01	180	HSS
	1.121	0.02	180	HSS
	1.5032	0.05	180	HSS

0 and 1. The variable scale data is one of the difficulties in the analysis, a numerical variable whose range of values is between 0 and 1000 is more voluminous in the analysis than a variable whose values are between 0 and 1 which would cause a problem of bias later. To do this, each value is normalized as follows:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \#(5)$$

This procedure only uses the min and max to normalize.

Then, these samples were converted into several series of images by a 108mp camera to obtain a database of images with an optimal resolution so that the prints left by the tools are recognizable. Each one is labeled with its Ra value and the parameters of the cutting conditions to establish the link between the two sources. Each combination of the cutting condition parameters in Tables 3 and 4 corresponds to a milled surface. We can see in Fig. 8 above, the overview of some machined plates.

4.3. Alignment

We go through a process of aligning our multimodal data, this process consists in linking the corresponding features of each modality. In our case, this process is done first by determining the features of the tool impressions left on the image during machining for machining tool recognition and the features extracted from the digital data. Finally, we identify the direct relationships that connect the features from these two data sources (see Fig. 9). On one side the numerical data and on the other side the image of the machined surfaces, the purpose of this process is to link each image to its roughness value and its cutting conditions. We can see this process in table x the link between each image and the cutting conditions of machining (roughness, feed per tooth, cutting speed, and cutting tools) forming a vector for each image in a CSV file (see Table 4).

4.4. Obtained results

We first used a dataset containing a single input (images) and the cutting condition parameters (cutting speed, feed rate, depth of cut) that will be the values to predict. For this, we implement a CNN architecture containing convolution and max pooling layers whose roles were in Section 3. This CNN is created for regression prediction by replacing the fully connected layer used for classification with the activation function by three other layers fully connected to the nodes representing each value to be predicted. These fully connected layers are given a linear activation function for regressing our cut condition parameters (see Fig. 10). We then train the CNN model with a mean square error (mse) loss function to predict continuous values.

We can use the learning curve to evaluate the behavior of machine learning models. In our case, the model has been trained on 60 epochs with a linear activation function and an MSE loss function. In Fig. 11, we can notice on the left the curve that represents the dynamics of the accuracy of the training and validation data which is going quite well despite the existence of a gap between the two databases. On the right, we have the loss curve on which we identify that the validation loss is higher than the training loss. In this case, it indicates that the training dataset is easier to predict for the model than the validation dataset. This

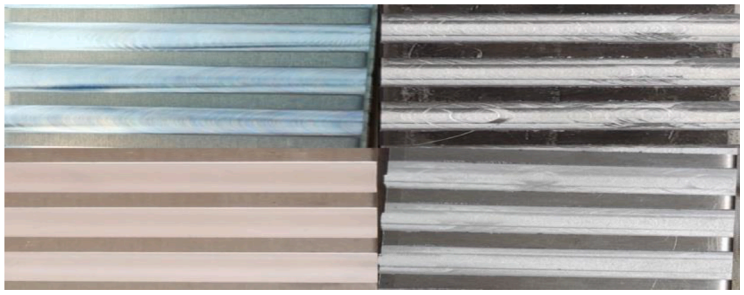

Fig. 8. Images of the machining on the aluminum plate.

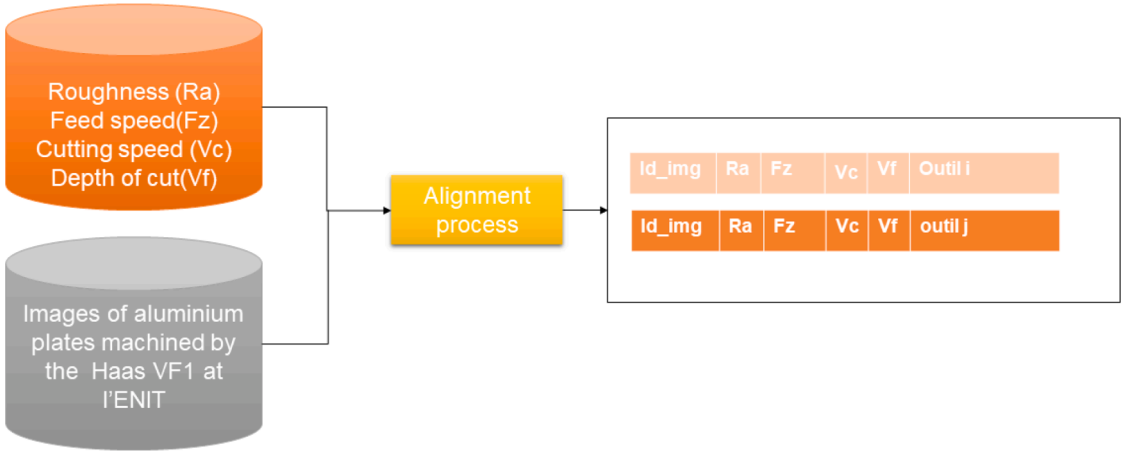

Fig. 9. : Alignment of multimodal data.

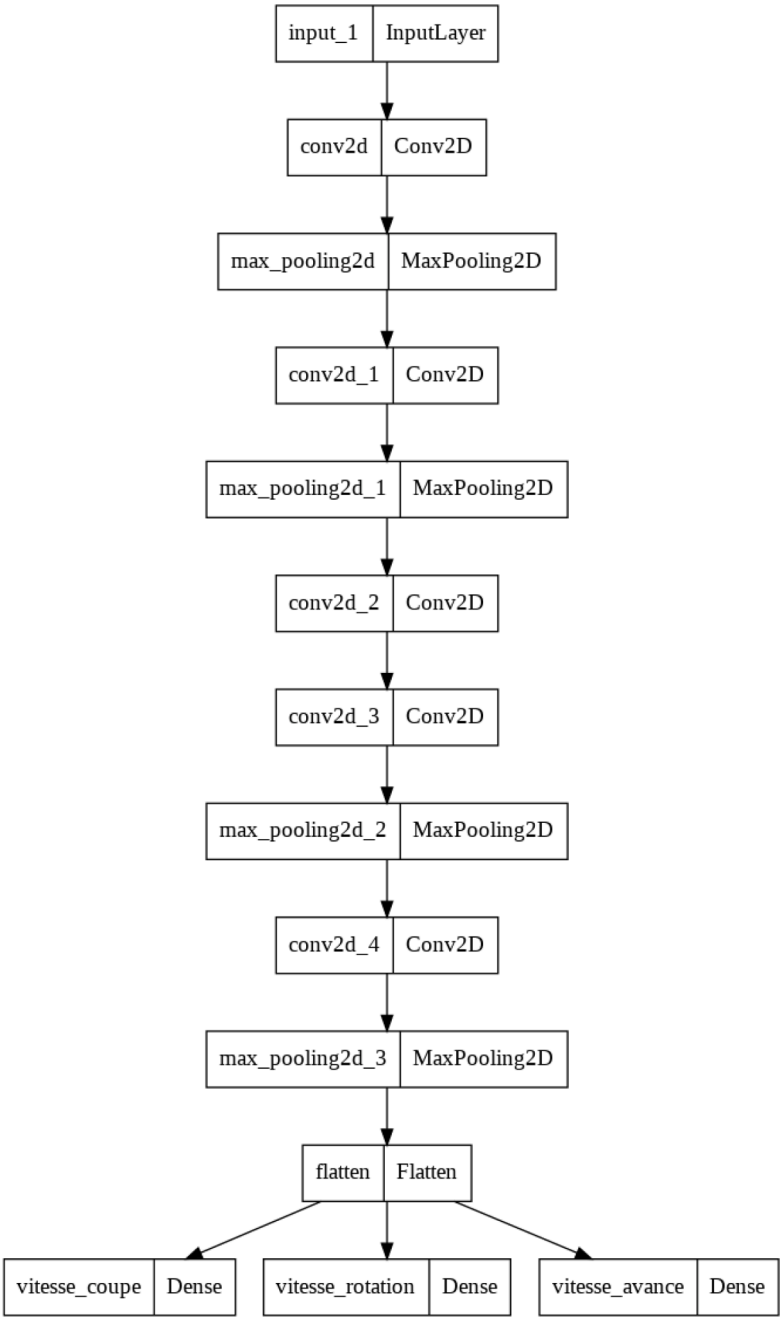

Fig. 10. The detailed architecture of the unimodal regression.

shows that the generalization ability of the unimodal model is weak and unreliable.

To correct this problem due to overfitting, we applied an approach called Early stopping a method that regularities deep learning models by stopping training after a certain number (wait) of epochs once the model's performance no longer progresses on a selected validation dataset (Naushad et al., 2021). In fact, throughout the training, the best weights of the model are kept and updated. In this work, the early stop was designed to monitor the validation loss curve and stop training when this validation loss no longer progresses for 4 epochs (the wait of 4). During the training, the best model weight is retained as soon as the value of the validation loss decreases. In this configuration, the training of the model stops after 17 epochs out of the 60 epochs initially planned (Fig. 12).

In the "Training and validation accuracy" section of the LSTM-CNN fusion model, the blue curve represents the accuracy of the training data, and the orange curve represents the accuracy of the test data (see Fig. 13). During the training of a deep model (or deep neural network), the model can be evaluated at each iteration (epoch).

During the training of a deep model (or deep neural network), the model can be evaluated at each iteration (epoch).

On the one hand, it is evaluated on the training dataset to give an idea of how well the model is learning, i.e., how well the model is learning.

On the other hand, the model can also be evaluated on the validation dataset which gives an idea of its generalization capacity (Kawaguchi et al., 2017). In our work, the samples in the validation dataset are not part of our model training data, i.e., the two datasets are distinct. The shape of the loss curves shows an inflection point in the validation loss which may be the point at which the training could be stopped, as

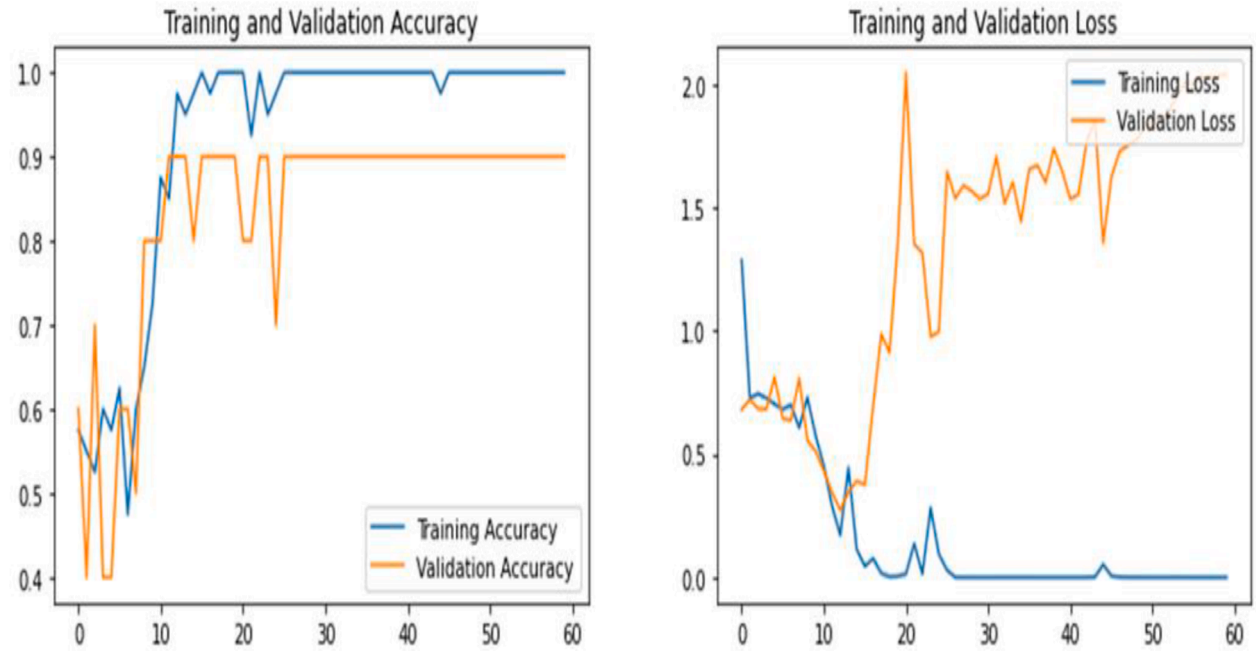

Fig. 11. The results obtained when training CNN unimodal models.

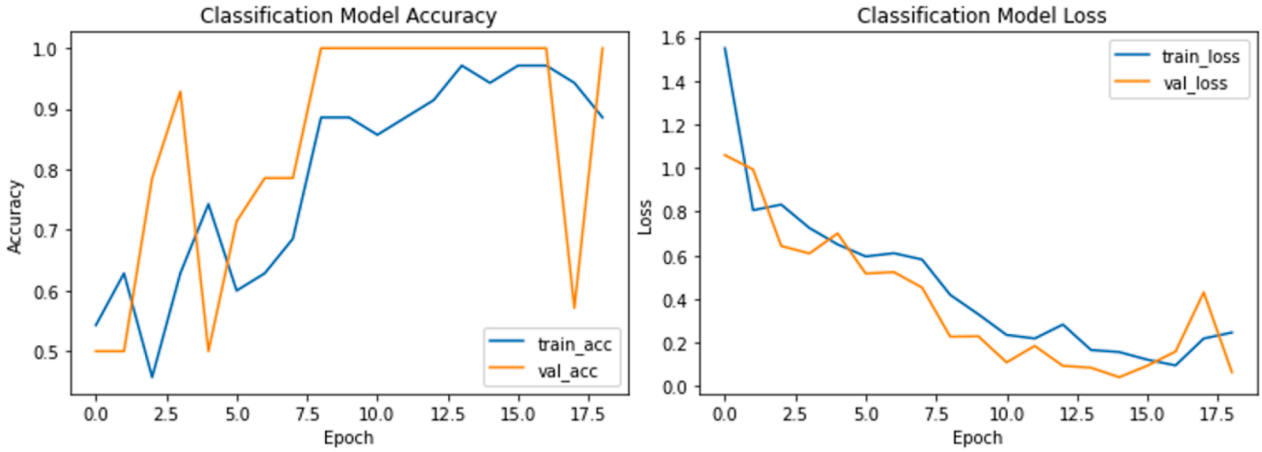

Fig. 12. Early stopping: training is stopped at epoch 17 when the performance of the validation loss stops decreasing even if the training loss is decreasing.

experimentation after this point will probably lead to an over-fitting of the model. When reading the loss curves in the validation phase, we observe the stability of the model until the 13th epoch which is the inflection point.

We notice that the precision continues to improve on both sets (training and testing) which allows us to avoid overfitting the data. In the second part, "Training and validation loss", the training and validation loss curves simultaneously experience a decrease, a sawtooth phase, and a stability phase.

We can observe in Table 5 that the multimodal fusion model outperforms the unimodal model with its ability to fit into the training and validation datasets. Thus, we justify our argument that the multimodal choice is better than the unimodal model. Multimodal systems capture more information than unimodal systems. Although unimodal models perform well, they do not consider information between modalities. Extracting important characteristics from a single modality before fusion does not consider the coherence and complementarity of the multimodal interaction and influences the final decision-making. In other words, one modality may provide additional information to the other modality, and the fusion features of the two make different contributions to the final decision. Because of the very limited amount of training samples in our dataset, the CNN model on the image dataset encounters overfitting phases on the validation dataset and the LSTM model on the numerical dataset fails to converge on both training and validation sets but the multimodal fusion model obtains better performance, which indicates its better generalization ability.

To show the effectiveness of the training, first, the distribution of the dataset changed from training (70%) and validation (30%) sets to training (50%), validation (20%), and test (30%) sets. Then we assigned an output layer to each parameter to see Fig. 6. The prediction of the parameters that have a numerical value is performed by a regression technique using the mean square error (MSE) (equation 5), and the mean absolute error (MAE) loss function as metrics (equation 6). Only the cutting tool is predicted with the softmax activation function (HSS and carbide), categorical cross-entropy loss function, and trueness as the metric. The curves in Fig. 14 show the deviation between the training set and the validation set over 16 epochs with a batch size of 16 using the MAE metric on the Vc and Fz outputs. This gap is calculated by the MAE loss function typically used for regression.

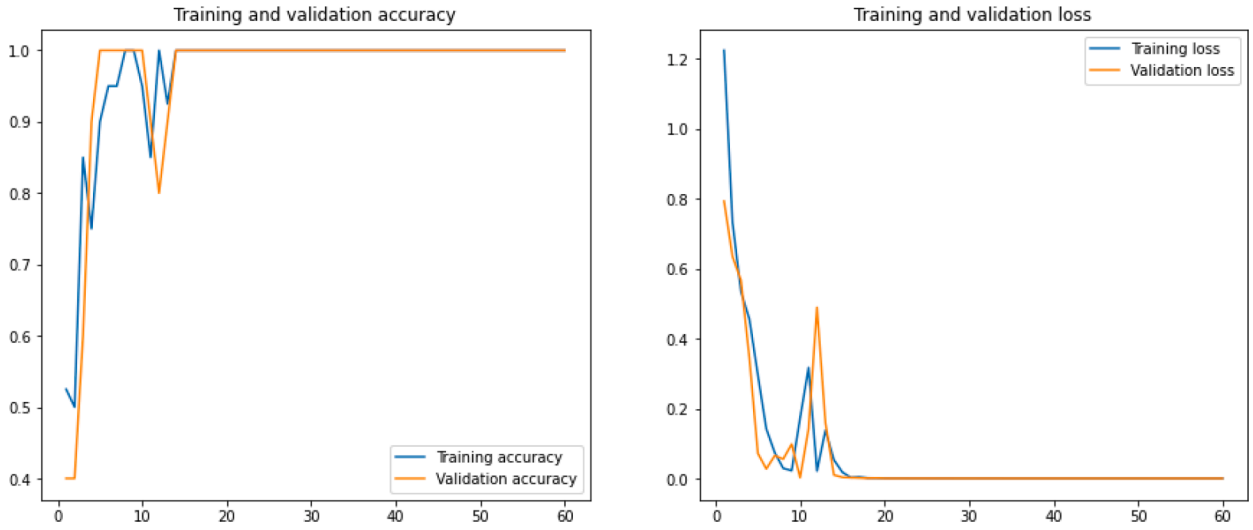

Fig. 13. The results obtained when training the LSTM and CNN models.

Table 5
Comparison between the unimodal models and the multimodal model.

Type of data	Model	Training loss rate	Training accuracy	Validation loss rate	Validation accuracy
Digital data	LSTM	0.7281	0.4701	0.6971	0.4667
Images data	CNN	0.0013	0.9667	1.8527	0.9000
Digital -Image data	CNN-LSTM	0.0068	0.9985	0.0196	0.9931

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad \#(5)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \#(6)$$

We notice on the figure that the MAE function performs more on the Vc output than Fz. At the beginning of the training, we find a big gap between the two sets but from 10 epochs on the two sets get closer.

Fig. 15 shows the training and validation curves plotted by the MSE function for Fz and Vc.

Fig. 15 shows a decrease in the error rates of the Fz and Vc outputs compared to the Mae curves. The MSE function places great emphasis on large errors, in contrast to MAE which is not sensitive to outliers.

Fig. 16 shows the loos and accuracy curves of the cutting tool output. The model manages to classify the cutting tool with an accuracy of over

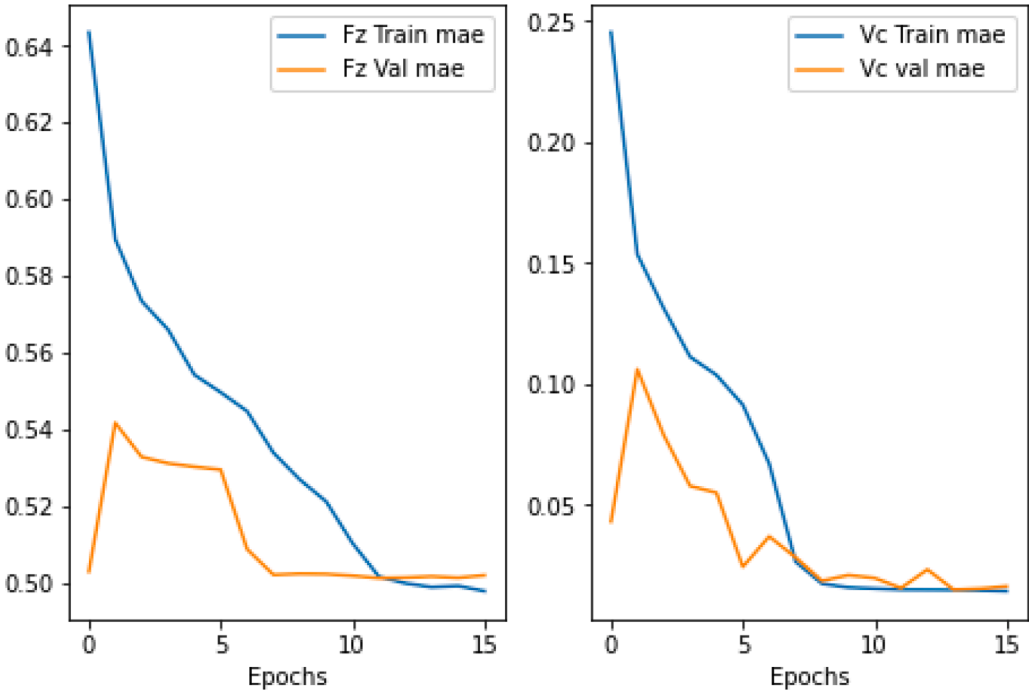

Fig. 14. Training and validation curves on Fz and Vc using the mae evaluator.

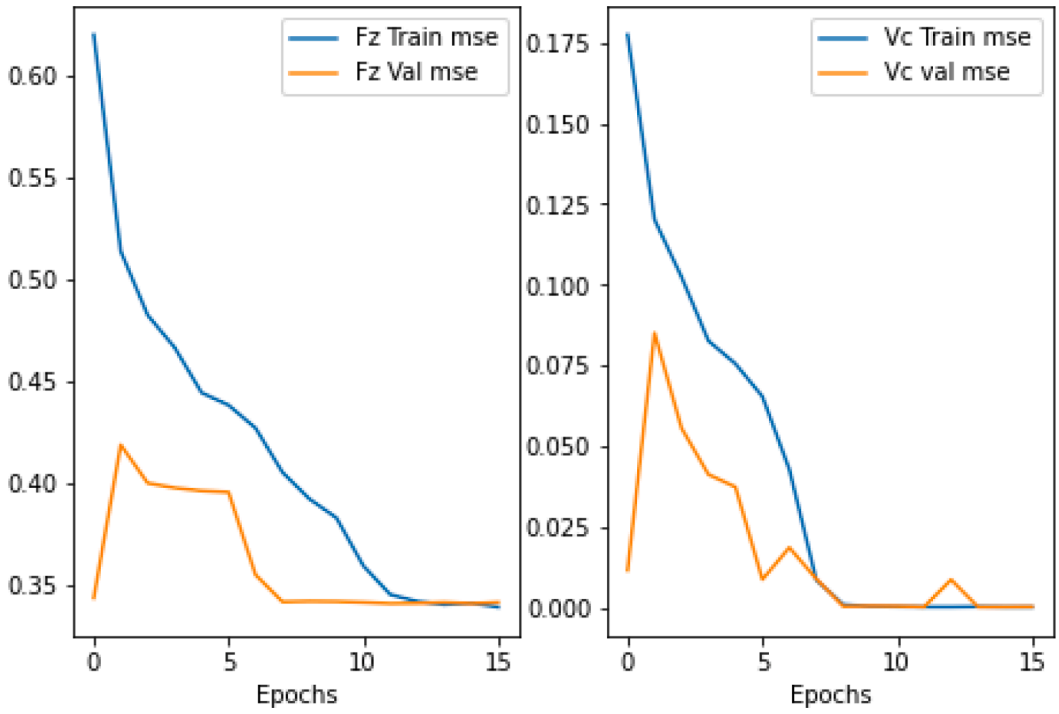

Fig. 15. Training and validation curves on Fz and Vc using the mse evaluator.

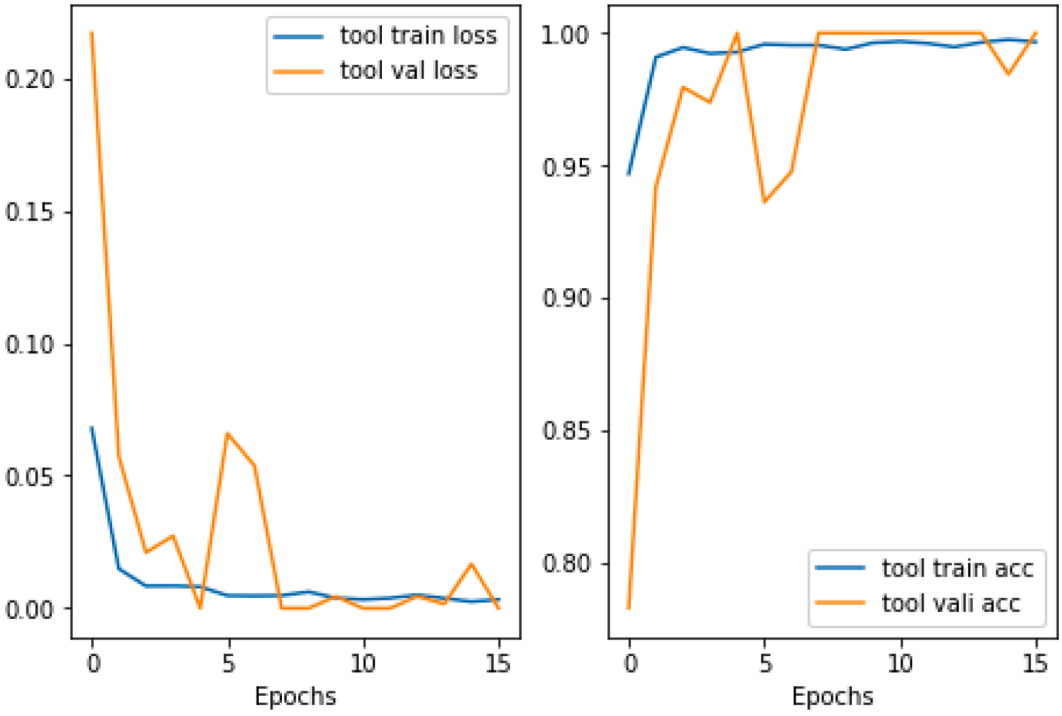

Fig. 16. Training validation loss and accuracy curves for tool classification.

Table 6
Metrics for evaluating regression outputs.

Outputs	Loss	MSE	MAE	Accuracy
Vc	0.013	0.00027	0.013	-
Fz	0.34	0.34	0.50	-
Tool	0.017	-	-	0.97

97% and a very low loss rate, the values of this step are summarized in Table 6.

Table 6 shows the values of the evaluation metrics on the test set. The metrics used for the test are MSE and MAE. The MAE metric is the most efficient in the training validation and test phases thanks to its ability not to be sensitive to outliers. MAE allows quantifying the error made by the model. The higher it is, the less efficient the model is. Contrary to MAE, MSE penalizes large errors more strongly than small errors.

Table 7
Classification report.

	précision	recall	f1-score
Carbide	1.00	0.90	0.95
HSS	0.97	1.00	0.99

For the tool prediction that was considered as a classification case (HSS and carbide), Table 7 shows the classification report. This report is used to show the main classification metrics per class which are precision, recall, and f1 score. These metrics are calculated concerning true and false positives and true and false negatives.

Precision is the ability of a model to not classify a sample from the HSS positive class that belongs to the Carbide class and is defined in equation 7.

$$Precision = \frac{True\ Positives}{(TruePositives + False\ Positives)} \#(7)$$

The recall is the ability of a model to search for all positive instances. It is defined in equation 8.

$$Recall = \frac{True\ Positives}{(False\ Positives + False\ Negatives)} \#(8)$$

The F1-score represents the percentage of correct predictions, equation 9.

$$F1 - score = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \#(9)$$

For further explanation of Table 7, Fig. 17 represents the confusion matrix of the cutting tool prediction. On the first-row notice that out of 727 samples of the carbide class 90 are classified in the HSS class and 2777 samples of the HSS class are all classified in their real class, which justifies a precision of 100% for the carbide class, and 97% for HSS.

5. Discussion

The methodology used was based on a late fusion strategy, as it allows for the best use of each of the models used. Sequencing the learning results of these models as late as possible promotes flexibility in managing multimodality. Some work in the literature often uses synthetic databases to model industrial conditions requiring multimodal fusion. In our case study, we worked on building a database populated by experiments performed on a CNC milling machine under the supervision of a domain expert. After running our neural network models on this data, the fusion model can perform classification tasks without the intervention of the expert to predict the choice of cutting parameters for the reproduction of a new piece. The resulting practical benefit is to facilitate the reproduction of machined pieces. To do this, it is first necessary to identify the choice of cutting parameters and the appropriate cutting tool on a CNC milling machine.

It is advisable to use several modalities (in this case, digital and images) to provide additional information about the considered machining process to obtain better performance than unimodal models. By applying late multimodal fusion to our neural network models, the accuracy increased significantly more than that found in the model using a single modality. Late multimodal fusion has proven to be particularly relevant when using digital data and images of the workpiece for cutting condition prediction and machining tool selection.

We have also adapted the structure of the model to our input data, and in particular to our desired output data. This also results in the ability to maintain and update the developed models when appropriate adjustments are warranted to improve the learning process. Estimating hyperparameters is an essential part of achieving better model quality but is extremely resource intensive. It may be useful to use the GridSearchCV method, which performs a deep search of the specified parameter values for an estimator, but it requires enormous computational power, to access several optimizers and choose the best one. There is a need to find appropriate ways to identify and implement efficiency measures to optimize the parameters as quickly as possible. This can be accomplished by trying fewer parameter options in each training run or by using RandomizedSearchCV (instead of GridSearchCV) which allows us to sample a fixed number of parameters from the specified distributions.

6. Conclusion and perspectives

In this paper, a major interest in the proposed methodology is the possibility to reuse knowledge from deep learning models. In terms of multimodality, exploitation is made of multimodal data whose intelligent combination within several models favors the generalization of the learning of our data on the training base and on the validation base for the prediction of the choice of cutting parameters. We use knowledge from the automatic machining domain to design a two-branch neural network model based on multimodal fusion for the prediction of the choice of cutting parameters and the appropriate cutting tool for machining. The cutting tool and cutting parameters are essential

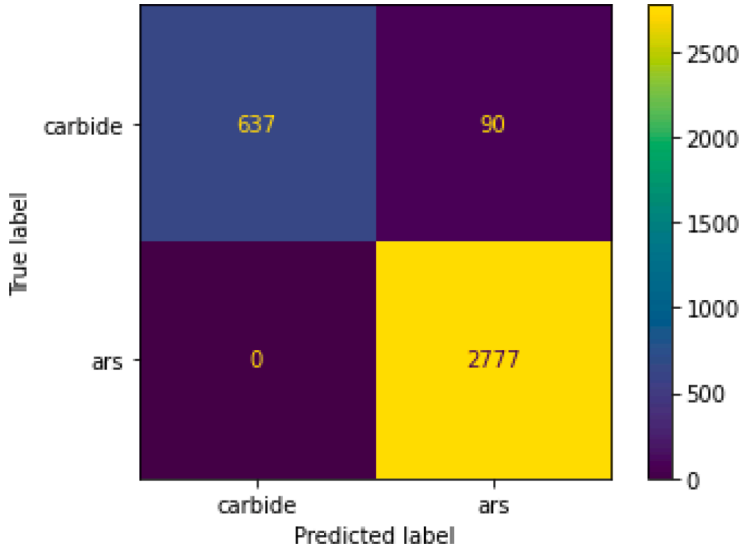

Fig. 17. Confusion matrix for tool prediction.

elements that influence the quality of the pieces to be machined. This quality plays a key role in the acceptance criteria of the high-precision components that have been machined.

Correct selection of the appropriate cutting tool and cutting parameters can significantly reduce the production time of the machining process and, consequently, the energy consumption. In circumstances where data quality is limited, the use of multimodality can allow the system to function even in the absence of one of the modalities. It is also observed that the performance of the multimodal fusion model is better than that of the unimodal model in predicting the cutting parameters. Theoretically, the contribution focuses on the deployment of an efficient procedure for multimodal machine learning. In particular, an alignment principle has been implemented to establish correspondences between images and numerical data. This alignment had an impact on the accuracy of the obtained results which are improved considerably. At the application level, the contribution concerns the development of an artificial intelligence approach for the prediction of parameters adapted to a type of machining. In this case, the proposed prediction aims at assisting in the choices guiding the appropriate configuration of tools and section parameters. We plan to perform several rounds of experiments by adding other types of cutting tools, such as drilling and turning tools, to increase the size of the database. Then we can use sound sensors to get vibration signals to make a vibration analysis which will allow us to raise and solve several problems like tool wear and machining fault detection. In addition, we consider applying another of the challenges of multimodality, co-learning, which allows the absence of one of the modalities to be handled in the test phase. In co-learning there are several cases of use of the modalities as in the case where all the data sources are used to perform the training and allow the absence of one of the modalities to perform the test. It is also possible that one or more modalities are absent during training but are found in the test phase. Co-learning is the transfer of knowledge between the modality, which will allow them to be complementary for the implementation of a multimodal system that works in the absence of one or more modalities.

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

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

References

Al-Dulaimi, A., Zabihi, S., Asif, A., & Mohammadi, A. (2019). A multimodal and hybrid deep neural network model for remaining useful life estimation. *Computers in Industry, 108*, 186–196. https://doi.org/10.1016/j.compind.2019.02.004

Alkhalaf, S. (2021). A robust variance information fusion technique for real-time autonomous navigation systems. *Measurement, 179*, Article 109441. https://doi.org/10.1016/j.measurement.2021.109441

AlZubi, A. A., Abugabah, A., Al-Maitah, M., & Ibrahim AlZobi, F. (2021). DL Multi-sensor information fusion service selective information scheme for improving the Internet of Things based user responses. *Measurement, 185*, Article 110008. https://doi.org/10.1016/j.measurement.2021.110008

Baltrušaitis, T., Ahuja, C., Morency, L.-P., 2017. Multimodal Machine Learning: A Survey and Taxonomy. ArXiv170509406 Cs.

Bayoudh, K., Knani, R., Hamdaoui, F., & Mtibaa, A. (2021). A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets. *Visual Computer*. https://doi.org/10.1007/s00371-021-02166-7

Bokade, R., Navato, A., Ouyang, R., Jin, X., Chou, C.-A., Ostadabbas, S., & Mueller, A. V. (2021). A cross-disciplinary comparison of multimodal data fusion approaches and applications: Accelerating learning through trans-disciplinary information sharing. *Expert Systems with Applications, 165*, Article 113885. https://doi.org/10.1016/j.eswa.2020.113885

Bouwmans, T., Javed, S., Sultana, M., & Jung, S. K. (2019). Deep neural network concepts for background subtraction: A systematic review and comparative evaluation. *Neural Networks, 117*, 8–66. https://doi.org/10.1016/j.neunet.2019.04.024

Charalampous, P. (2021). Prediction of cutting forces in milling using machine learning algorithms and finite element analysis. *Journal of Materials Engineering and Performance*. https://doi.org/10.1007/s11665-021-05507-8

Che, C., Wang, H., Ni, X., & Lin, R. (2020). Hybrid multimodal fusion with deep learning for rolling bearing fault diagnosis. *Measurement, 108655*. https://doi.org/10.1016/j.measurement.2020.108655

Chen, X., Li, C., Tang, Y., Li, L., & Li, H. (2021). Energy efficient cutting parameter optimization. *Frontiers in Mechanical Engineering, 16*, 221–248. https://doi.org/10.1007/s11465-020-0627-x

Chen, Xue, Zhang, L., Liu, T., & Kamruzzaman, M. M. (2019). Research on deep learning in the field of mechanical equipment fault diagnosis image quality. *Journal of Visual Communication and Image Representation, 62*, 402–409. https://doi.org/10.1016/j.jvcir.2019.06.007

Cuayáhuitl, H. (2020). A data-efficient deep learning approach for deployable multimodal social robots. *Neurocomputing, 396*, 587–598. https://doi.org/10.1016/j.neucom.2018.09.104

Duan, L., Xie, M., Wang, J., & Bai, T. (2018). Deep learning enabled intelligent fault diagnosis: Overview and applications. *Journal of Intelligent and Fuzzy Systems, 35*, 5771–5784. https://doi.org/10.3233/JIFS-17938

Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., & Dehmer, M. (2020). An introductory review of deep learning for prediction models with big data. *Frontiers in Artificial Intelligence, 3*. https://doi.org/10.3389/frai.2020.00004

Fu, P., Wang, J., Zhang, X., Zhang, L., & Gao, R. X. (2020). Dynamic routing-based multimodal neural network for multi-sensory fault diagnosis of induction motor. *Journal of Manufacturing Systems, 55*, 264–272. https://doi.org/10.1016/j.jmsy.2020.04.009

Gupta, A., Anpalagan, A., Guan, L., & Khwaja, A. S. (2021). Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. *Array, 100057*. https://doi.org/10.1016/j.array.2021.100057

Gürdür Broo, D., Boman, U., & Törngren, M. (2021). Cyber-physical systems research and education in 2030: Scenarios and strategies. *Journal of Industrial Information Integration, 21*, Article 100192. https://doi.org/10.1016/j.jii.2020.100192

Graves, A., Jaitly, N., Mohamed, A., 2013. Hybrid speech recognition with deep bidirectional LSTM, in: Proceedings of the IEEE workshop on automatic speech recognition and understanding. Presented at the 2013 IEEE workshop on automatic speech recognition and understanding, pp. 273–278. 10.1109/ASRU.2013.6707742.

Hatami, N., Cho, T.-H., Mechtouff, L., Eker, O.F., Rousseau, D., Frindel, C., 2022. CNN-LSTM based multimodal MRI and clinical data fusion for predicting functional outcome in stroke patients. 10.48550/ARXIV.2205.05545.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hu, L., Tang, R., Cai, W., Feng, Y., & Ma, X. (2019). Optimisation of cutting parameters for improving energy efficiency in machining process. *Robotics and Computer-Integrated Manufacturing, 59*, 406–416. https://doi.org/10.1016/j.rcim.2019.04.015

de la Peña Zarzuelo, I., Soeane, M. J. F., & López Bermúdez, B. (2020). Industry 4.0 in the port and maritime industry: A literature review. *Journal of Industrial Information Integration, 20*, Article 100173. https://doi.org/10.1016/j.jii.2020.100173

Indolia, S., Goswami, A. K., Mishra, S. P., & Asopa, P. (2018). Conceptual understanding of convolutional neural network- A deep learning approach. In *, 132. Procedia Comput. Sci., International Conference on Computational Intelligence and Data Science, April 7th to 8th, 2018* (pp. 679–688). Gurugram, India: The NorthCap University. https://doi.org/10.1016/j.procs.2018.05.069.

Kawaguchi, K., Kaelbling, L.P., Bengio, Y., 2017. Generalization in Deep Learning. https://doi.org/10.48550/ARXIV.1710.05468.

Kim, T., & Kim, H. Y. (2019). Forecasting stock prices with a feature fusion LSTM-CNN model using different representations of the same data. *Plos One, 14*, Article e0212320. https://doi.org/10.1371/journal.pone.0212320

Kumar, S., Kolekar, T., Patil, S., Bongale, A., Kotecha, K., Zaguia, A., & Prakash, C. (2022). A low-cost multi-sensor data acquisition system for fault detection in fused deposition modelling. *Sensors, 22*, 517. https://doi.org/10.3390/s22020517

Li, Z., Liu, X., Incecik, A., Gupta, M. K., Królczyk, G. M., & Gardoni, P. (2022). A novel ensemble deep learning model for cutting tool wear monitoring using audio sensors. *Journal of Manufacturing Processes, 79*, 233–249. https://doi.org/10.1016/j.jmapro.2022.04.066

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42*, 60–88. https://doi.org/10.1016/j.media.2017.07.005

Liu, H., Fang, T., Zhou, T., & Wang, L. (2018). Towards robust human-robot collaborative manufacturing: Multimodal fusion. *IEEE Access, 6*, 74762–74771. https://doi.org/10.1109/ACCESS.2018.2884793

Lu, J., Behbood, V., Hao, P., Zuo, H., Xue, S., & Zhang, G. (2015). Transfer learning using computational intelligence: A survey. *Knowledge-Based Systems, 80*, 14–23. https://doi.org/10.1016/j.knosys.2015.01.010, 25th anniversary of Knowledge-Based Systems.

Luo, J., & Zhang, X. (2022). Convolutional neural network based on attention mechanism and Bi-LSTM for bearing remaining life prediction. *Applied Intelligence, 52*, 1076–1091. https://doi.org/10.1007/s10489-021-02503-2

Ma, M., Sun, C., & Chen, X. (2018). Deep coupling autoencoder for fault diagnosis with multimodal sensory data. *IEEE Transactions on Industrial Informatics, 14*, 1137–1145. https://doi.org/10.1109/TII.2018.2793246

Mou, L., Zhou, C., Zhao, P., Nakisa, B., Rastgoo, M. N., Jain, R., & Gao, W. (2021). Driver stress detection via multimodal fusion using attention-based CNN-LSTM. *Expert Systems with Applications, 173*, Article 114693. https://doi.org/10.1016/j.eswa.2021.114693

Naushad, R., Kaur, T., & Ghaderpour, E. (2021). Deep transfer learning for land use and land cover classification: A comparative study. *Sensors, 21*, 8083. https://doi.org/10.3390/s21238083

Okokpujie, I. P., Ohunakin, O. S., Bolu, C. A., & Okokpujie, K. O. (2018). Experimental data-set for prediction of tool wear during turning of Al-1061 alloy by high speed steel cutting tools. *Data Brief, 18*, 1196–1203. https://doi.org/10.1016/j.dib.2018.04.003

Palade, V., Wermter, S., Ruiz-Garcia, A., Braga, A. D. P., & Took, C. C. (2021). Guest Editorial: Special issue on deep representation and transfer learning for smart and connected health. *IEEE Transactions on Neural Networks and Learning Systems, 32*, 464–465. https://doi.org/10.1109/TNNLS.2021.3049931

Poria, S., Cambria, E., Howard, N., Huang, G.-B., & Hussain, A. (2016). Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing, 174*, 50–59. https://doi.org/10.1016/j.neucom.2015.01.095

Qin, Y., Xiang, S., Chai, Y., & Chen, H. (2020). Macroscopic–microscopic attention in LSTM networks based on fusion features for gear remaining life prediction. *IEEE Transactions on Industrial Electronics, 67*, 10865–10875. https://doi.org/10.1109/TIE.2019.2959492

Rahate, A., Mandaokar, S., Chandel, P., Walambe, R., Ramanna, S., & Kotecha, K. (2022). Employing multimodal co-learning to evaluate the robustness of sensor fusion for industry 5.0 tasks. *Soft Computing*. https://doi.org/10.1007/s00500-022-06802-9

Shi, J., Peng, D., Peng, Z., Zhang, Z., Goebel, K., & Wu, D. (2022). Planetary gearbox fault diagnosis using bidirectional-convolutional LSTM networks. *Mechanical Systems and Signal Processing, 162*, Article 107996. https://doi.org/10.1016/j.ymssp.2021.107996

Siyu, L., Shaoluo, H., Yangyang, Z., Lijun, C., & Weiyi, W. (2020). Deep learning in fault diagnosis of complex mechanical equipment. *International Journal of Performability Engineering, 16*, 1548. https://doi.org/10.23940/ijpe.20.10.p6.15481555

Wang, J., Ma, Y., Zhang, L., Gao, R. X., & Wu, D. (2018). Deep learning for smart manufacturing: Methods and applications. *Journal of Manufacturing Systems, 48*, 144–156. https://doi.org/10.1016/j.jmsy.2018.01.003. Special Issue on Smart Manufacturing.

Wen, T.-H., Gasic, M., Mrksic, N., Su, P.-H., Vandyke, D., Young, S., 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. ArXiv150801745 Cs.

Xiang, L., Yang, X., Hu, A., Su, H., & Wang, P. (2022). Condition monitoring and anomaly detection of wind turbine based on cascaded and bidirectional deep learning networks. *Applied Energy, 305*, Article 117925. https://doi.org/10.1016/j.apenergy.2021.117925

Yan, H., Wan, J., Zhang, C., Tang, S., Hua, Q., & Wang, Z. (2018). Industrial big data analytics for prediction of remaining useful life based on deep learning. *IEEE Access, 6*, 17190–17197. https://doi.org/10.1109/ACCESS.2018.2809681

Yang, Z., Baraldi, P., & Zio, E. (2021). A multi-branch deep neural network model for failure prognostics based on multimodal data. *Journal of Manufacturing Systems, 59*, 42–50. https://doi.org/10.1016/j.jmsy.2021.01.007

Yu, Bihui, Wei, J., Yu, Bo, Cai, X., Wang, K., Sun, H., Bu, L., & Chen, X. (2022). Feature-guided multimodal sentiment analysis towards industry 4.0. *Computers and Electrical Engineering, 100*, Article 107961. https://doi.org/10.1016/j.compeleceng.2022.107961

Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2019). Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing, 115*, 213–237. https://doi.org/10.1016/j.ymssp.2018.05.050

Zhou, F., Hu, P., Yang, S., & Wen, C. (2018). A multimodal feature fusion-based deep learning method for online fault diagnosis of rotating machinery. *Sensors, 18*, 3521. https://doi.org/10.3390/s18103521