Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 00 (2023) 000–000

www.elsevier.com/locate/procedia

17th CIRP Conference on Intelligent Computation in Manufacturing Engineering

A meta-learning strategy based on deep ensemble learning for tool condition monitoring of machining processes

Jose Joaquin Peralta Abadia*, Mikel Cuesta Zabaljauregui, Felix Larrinaga Barrenechea

Mondragon Unibertsitatea, Loramendi Kalea, 4, 20500 Arrasate, Spain

* Corresponding author. *E-mail address:* jjperalta@mondragon.edu

Abstract

For Industry 4.0, tool condition monitoring (TCM) of machining processes aims to increase process efficiency and quality and lower tool maintenance costs. To this end, TCM systems monitor variables of interest, such as tool wear. In this paper, a novel meta-learning strategy based on ensemble learning and deep learning (DL) is proposed for tool wear monitoring and is compared with state-of-the-art DL models selected from recent literature, using open-access datasets as input validating its implementation in an industrial scenario. As a result of this study, a novel meta-learning strategy for tool wear monitoring with minimum error is proposed and validated.

Peer-review under responsibility of the scientific committee of the 17th CIRP Conference on Intelligent Computation in Manufacturing Engineering, 12-14 July, Gulf of Naples, Italy

Keywords: tool wear; deep learning; industry 4.0, tool condition monitoring; ensemble learning

1. Introduction

Machining processes, such as milling, are widely used in manufacturing to achieve highly accurate machine parts and good surface integrity [1]. To satisfy the quality requirements of the finished piece, tool condition monitoring (TCM) systems are required to improve product quality, process dependability, and production efficiency [2]. The primary aim of TCM is to identify the appropriate time to replace cutting tools. Changing tools too soon disrupts production times, and too late can cause damage to equipment, machines, and workpieces.

However, TCM of machining processes, and in particular deep learning (DL)-based TCM, is yet to fully reach the shop floor [2]. This is because DL models usually require big data for training, which is challenging in machining processes where data is generally not publicly available or is unlabelled [3].

Aiming to mitigate the problem of data availability, open-access datasets have been published in the literature, such as the NASA Ames/UC Berkeley milling dataset [4]. As a result, several authors have proposed DL models trained with this dataset.

Aghazadeh et al. (2018) implemented a convolutional neural network (CNN) model in combination with spectral subtraction of wavelet packets, using the current signals of the dataset, achieving a root-mean-squared error (RMSE) of 0.088 mm [5]. More recently, Cai et al. (2020) presented a hybrid model based on long short-term memory (LSTM) networks. The model was trained with all signals and cutting conditions of the dataset, using 4 cases for testing and the remaining 12 for training and achieving a RMSE of 0.0456 mm. The LSTM layer was used for temporal encoding of features, and thereafter, a non-linear regression network combined the temporal features obtained from the LSTM with the cutting conditions to perform the predictions [6]. Another hybrid LSTM model, comprised of bidirectional LSTM and encoder-decoder LSTM layers, was proposed by Kumar et al. (2022). The model used time and frequency features extracted from the vibration signals of the dataset, achieving a RMSE of 0.0364 mm [7]. Finally, Pillai and

2212-8271
Peer-review under responsibility of the scientific committee of the 17th CIRP Conference on Intelligent Computation in Manufacturing Engineering, 12-14 July, Gulf of Naples, Italy

Vadakkepat (2022) presented a temporal multivariate 3D convolutional network model, trained with 3D features from the signals obtained from kernel-based transformations, and achieving a RMSE of 0.0424 mm [8].

Although good performance has been obtained from the models in scientific experiments, the margin of error is still not acceptable for industrial implementations. Furthermore, the selection of signals from the dataset to be used as inputs requires a systematic approach. Therefore, the potential of applying DL to TCM requires further research, such as meta-learning strategies that combine DL models with ensemble learning techniques [9].

In this paper, a novel meta-learning strategy based on deep ensemble learning (DEL) is proposed for tool wear monitoring. The strategy was compared with state-of-the-art DL models selected from recent literature, using the NASA Ames/UC Berkeley open-access dataset as input. As such, the contribution of this paper is twofold:

- Meta-learning strategy: A novel meta-learning strategy based on deep ensemble learning (DEL) compared against state-of-the-art DL models selected from recent literature, proving a superior prediction performance.
- An analysis of the signals from the NASA Ames/UC Berkeley dataset to identify ideal signals to be used as inputs for DL learning models. The signals are analysed, cleaned, and augmented. Then, four combinations of signals (all signals, current and acoustic emission signals, current signal, and vibration signal) are compared in relation to their effect on DL model performance.

The reminder of this paper is structured as follows. Section 2 describes the open-access dataset used in this study. Section 3 describes the methodology followed to implement the meta-learning strategy. Thereafter, Section 4 presents results and discussion. Finally, Section 5 presents conclusions and outlook on future work.

2. Dataset description

The NASA Ames/UC Berkeley open-access dataset [4] was used in this study as input for training the meta-learning strategy based on DEL. The dataset encompasses 16 face milling experiments that were performed on a milling machine under varying cutting conditions. Three types of sensors, i.e., acoustic emission (AE) sensors, vibration sensors, and current sensors were employed to collect data with a sampling rate of 250 Hz. Specifically, the sensors collected signals including spindle motor current AC (smcAC), spindle motor current DC (smcDC), table vibration (vib_table), spindle vibration (vib_spindle), table AE (AE_table), and spindle AE (AE_spindle). In addition, the dataset was enriched with process information, such as case number, experimental run count, tool wear (VB), experiment duration, and cutting conditions. Cutting conditions included depth of cut (DOC), feed rate, and material type.

A total of 167 runs were performed for approximately 36 s each, containing 9000 measurement points per run. The number of runs per case varied according to the extent of VB assessed between runs at variable intervals. Specifically, VB was not recorded for all runs. Moreover, the degree of tool wear surpassed the manufacturer recommended VB limit in some cases.

The experimental conditions of the cases are presented in Table 1, and include two values for DOC (1.5 and 0.25 mm), two values for feed rate (0.5 and 0.25 mm/rev), and two material types (1-cast iron and 2-stainless steel). The cutting tools used were KC710 inserts, the cutting speed was 200 m/min (or 826 rev/min), and the workpiece size was 483 mm x 178 mm x 51 mm. Eight combinations of cutting conditions were defined, and each combination was repeated a second time with a new set of cutting tools.

Table 1. Experimental conditions of the NASA Ames/UC Berkeley dataset.

Case	DOC	Feed rate	Material	Case	DOC	Feed rate	Material
1	1.5	0.5	1	9	1.5	0.5	1
2	0.75	0.5	1	10	1.5	0.25	1
3	0.75	0.25	1	11	0.75	0.25	1
4	1.5	0.25	1	12	0.75	0.5	1
5	1.5	0.5	2	13	0.75	0.25	2
6	1.5	0.25	2	14	0.75	0.5	2
7	0.75	0.25	2	15	1.5	0.25	2
8	0.75	0.5	2	16	1.5	0.5	2

Pearson correlation coefficient was applied to the dataset, obtaining the correlations between the dataset features, and is depicted as a correlation matrix in Fig. 1. Of the signals, smcDC reported the highest correlation with VB, and also presented a high correlation with AE signals. Fig. 2 illustrates the VB histogram of the dataset, in which an exponential distribution is observed. VB progressed slowly in both the break-in and regular wear stages of the cutting tool. The VB curve increased exponentially in the high and critical wear stages, until the tool was no longer usable.

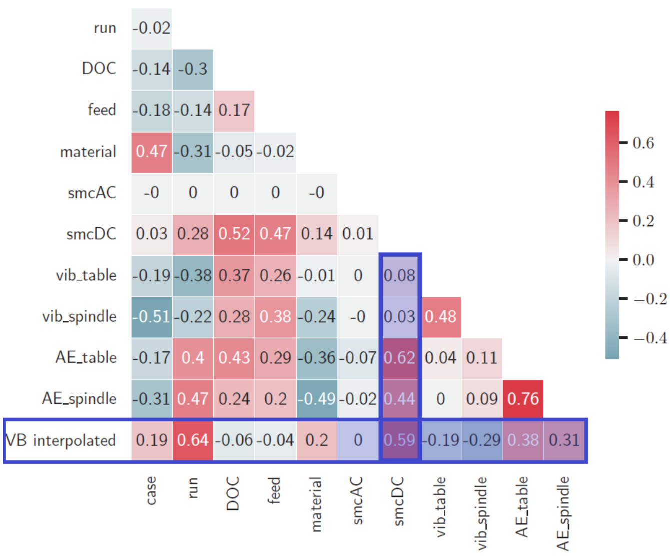

Fig. 1. Correlation matrix of the NASA Ames/UC Berkeley dataset.

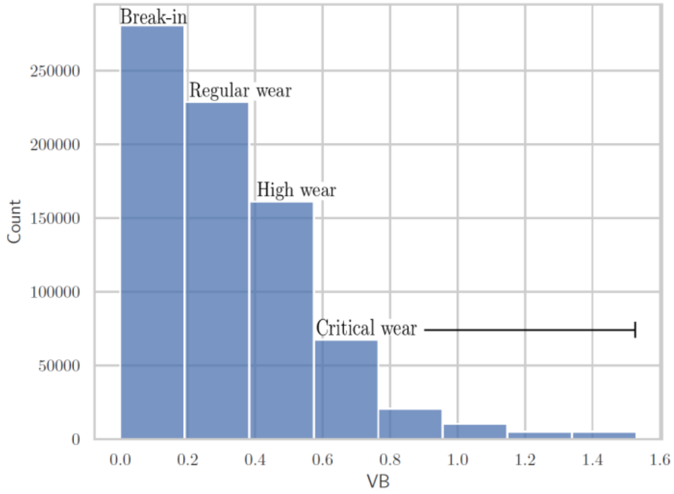

Fig. 2. VB histogram of the NASA Ames/UC Berkeley dataset.

3. Methodology

The methodology to achieve TCM using the novel meta-learning strategy was comprised of three steps. First, data pre-processing was performed on the NASA Ames/UC Berkeley open-access dataset to clean and prepare it for training the DEL model. Second, machine learning (ML) models were developed and implemented as baseline models. Finally, the meta-learning strategy based on DEL was developed and implemented.

3.1. Data pre-processing

The dataset containes measurements collected during entry, regular, and exit cuts of the experiments. In this study, the entry and exit cut portions of the signals were omitted, focusing only on the regular cut portion of the machining process. Furthermore, since some cases did not record VB, linear interpolation was performed to use all data available. Thereafter, signals for each run were evaluated. Data acquired in eight runs were corrupted or had undocumented events and were omitted in this study, resulting in 159 runs for training and testing. In addition, 22 runs had signals with noisy values, which could have a negative impact on the prediction capabilities of the meta-learning strategy. For predicting tool wear, the global behaviour of the signal is more important than localized events (e.g., chipping). Therefore, a moving average with size 20 was applied to average out the noisy values, while maintaining the global behaviour of the signals. The following are the two groups of runs that were treated:

- Omitted
  - Case 1 - Runs 16 and 17: VB lowers after run 15.
  - Case 2 - Run 5: Missing data in AE_table.
  - Case 2 - Run 6: Corrupt data in AE_spindle.
  - Case 7 - Run 4: Corrupt data in AE_table.
  - Case 8 - Run 3: Missing data in AE_table.
  - Case 12 - Run 1: Corrupt data in all signals
  - Case 12 - Run 12: Undocumented event in all signals.
- Noise
  - Case 3 - Run 9.
  - Case 7 - Run 8.
  - Case 8 - Run 4.
  - Case 10 - Runs 2 and 10.
  - Case 11 - Runs 10 and 21.
  - Case 12 - Runs 3 and 7.
  - Case 13 - Runs 3, 6, 8, 9, 13, and 14.
  - Case 14 - Runs 1, 2, 3, 6, and 10.
  - Case 15 - Runs 1, 2, 3, 4, 6, and 7.

3.2. Machine learning baseline models

Six ML models were trained with the input data as baseline models: (i) decision tree, (ii) random forest, (iii) support vector machine (SVM), (iv) gradient boosting, (v) XGBoost, and (vi) k-nearest neighbours (kNN). For the sake of brevity, detailed descriptions of the algorithms are omitted but can be found in [10,11].

The feature extraction methodology proposed in [12] was adopted to train the baseline ML models. Time domain, frequency domain, and time-frequency domain features were extracted, and are presented in Table 2, with a total of 54 extracted features. A more detailed description of the extracted features can be found in [12]. Afterwards, the features were normalized with z-normalization using the standard z-score, calculated as $z = (x - \mu) / \sigma$, where μ is the mean of the feature, x is the value of the feature, and σ is the standard deviation of the feature.

Table 2. Extracted features of the time, frequency, and time-frequency domains.

Domain	Feature
Time	RMS
	Variance
	Maximum
	Kurtosis
	Skewness
	Peak-to-peak
Frequency	Spectral skewness
	Spectral kurtosis
Time-frequency	Wavelet energy

Given the high quantity of features and the inherent high correlation among them, a dimensionality reduction approach was required. To this end, the principal component analysis (PCA) technique was used [13]. The variance of the dataset that each component represents was analysed to determine the number of principal components to be chosen. At least 95% of variance was considered to properly represent the dataset [12].

3.3. Meta-learning strategy based on deep ensemble learning

Ensemble learning trains multiple ML or DL models, called base learners, to output several weak predictions from the same problem. The predictions are generally combined using voting and averaging mechanisms, which results in better performance than those of the models by themselves [14]. Recently, meta-learning has been proposed for combining predictions, to improve the performance of ensemble learning. Meta-learning consists of learning from the outputs of each of the learners and

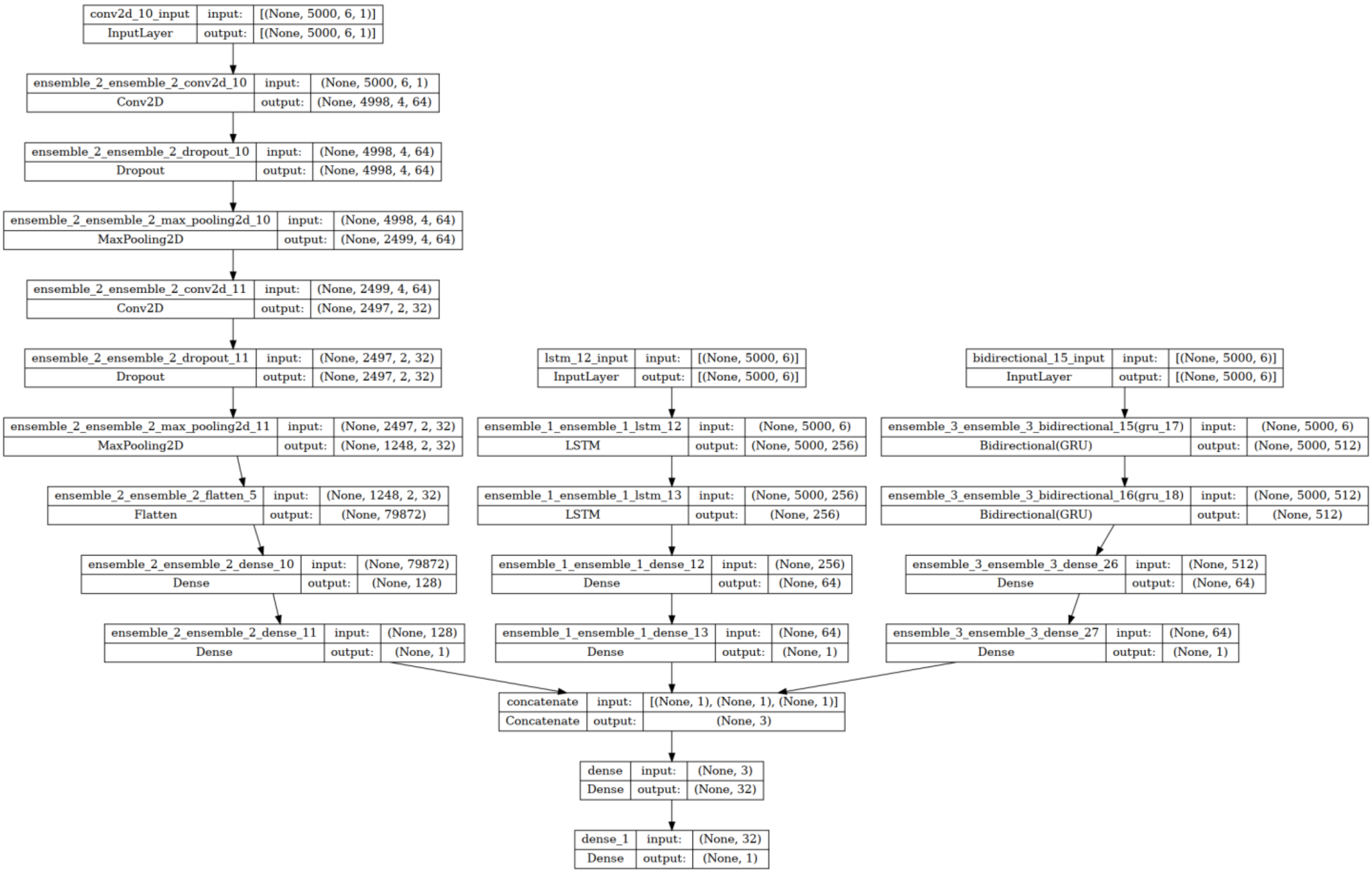

Fig. 3. Meta-learning model architecture.

making predictions based on the outputs combined. Hence, well performing base learners help offset those that perform badly for some problems, and vice versa for other problems. The most commonly used meta-learning strategy is stacked generalization (or stacking), which learns how to best combine the outputs of the base learners by using another ML or DL model [15].

A heterogeneous DEL approach was implemented, comprised of LSTM, bidirectional gated recurrent unit (BiGRU), and CNN models as base learners. Moreover, a deep neural network (DNN) was used as meta-learner, combining the predictions from the base learners.

First, the base learners were trained with the signals as input data. A DL stacking meta-learner was subsequently defined and trained, where the trained base learners were used as initial layers. As a result, the weak predictions were the input features of the meta-learner. Fig. 3 depicts the architecture of the meta-learning model.

The model was evaluated using all available signals, owing to the benefits of sensor fusion [2]. Moreover, other combinations were explored as well. Fig. 1 shows that smcDC had the highest correlation to VB, followed by both AE signals. In general, AE signals have high accuracy and resolution and have proven to be reliable for detecting events in machining processes [1]. Therefore, a combination of smcDC with AE_table and AE_spindle signals was explored to evaluate the performance of the approach with less signals but with a relatively high correlation among them. The performance of the approach was compared with state-of-the-art DL models selected from recent literature [5–8]. Since some of the DL models were trained only with either the vibration or the current signals, the use of smcDC as single input, as well as vib_spindle, were also explored for training the meta-learning strategy. Consequently, four strategies for training meta-learning models with varied input data were explored: (i) all sensor signals, (ii) AE_table, AE_spindle and smcDC sensor signals, (iii) smcDC sensor signal, and (iv) vib_spindle sensor signal.

4. Results and discussion

Six ML baseline models and a meta-learning model based on DEL were trained and tested. For the baseline models, time, frequency, and time-frequency domain features were extracted and *z*-normalized, for a total of 54 features (nine features per signal). PCA was selected for dimensionality reduction and the explained variance of the components is presented in Fig. 4.

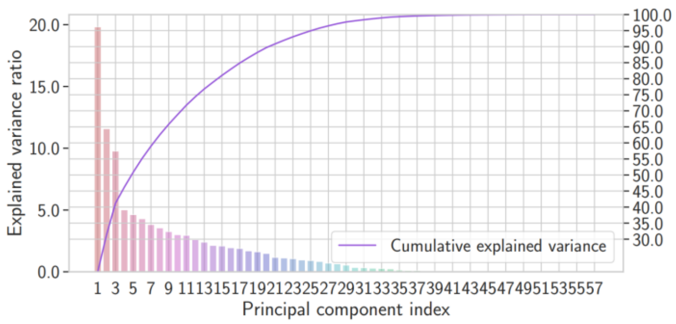

Fig. 4. Explained variance of PCA of the NASA Ames/UC Berkeley dataset.

The first 25 components were selected, as they represent 95% of the variance in the dataset.

Table 3 presents the hyperparameters of the baseline models, as well as performance metrics during testing. The hyperparameters were obtained using a randomized search cross validation method. Coefficient of determination (R^2), mean absolute error (MAE) and root-mean-squared error (RMSE) were chosen as performance metrics. The performance of the models is best when closest to one for R^2 and closest to zero for MAE and RMSE. The best performing model was kNN, followed closely by XGBoost. However, the scores indicate that the models could have an error in average of 0.0739 mm in its prediction. In industrial scenarios, a maximum tool wear of 0.3 mm is recommended by manufacturers. Thus, the error in the predictions represents a 25% of the industrial tool life and would not be acceptable in shop floors.

After training and testing the baseline models, the meta-learning model based on DEL was implemented. As with the baseline models, the data was *z*-normalized. Furthermore, since DL models require big data, a sliding window approach was adopted to augment the dataset. The sliding window was of size 250 (one second) and stride 25 (1/10 of a second), increasing the dataset size from 166 datapoints with a sequence length of 5400, to 31323 datapoints with a sequence length of 250.

All strategies shared the same model hyperparameters. The LSTM, BiGRU, and DEL models were trained for 1000 epochs, with an early stop after 50 epochs without model improvement. The CNN model required more epochs to generalize knowledge, so 4000 epochs with an early stop after 200 epochs were defined. All models used the ADAM optimizer with a learning rate of 0.0001 and RMSE as loss function. To avoid overfitting, a dropout of 10% and L2 regularization factor of 0.00001 were implemented. The dataset was split stochastically into 48% for training, 12% for validation, and 40% for testing. The split was made stochastically to account for the variability in cutting conditions and tools that may occur in industrial shop floors.

The performance results of the meta-learning model grouped by input data strategies, as well as a comparison with state-of-the-art DL models, is presented in Table 4. Results in the table prove the meta-learning strategy benefits, improving the quality of the predictions by combining the predictions of the base learners. For the base learners, the LSTM and BiGRU models performed better than the CNN model with combinations of signals. However, when using individual signals, the LSTM model was the worst predictor.

The model outperformed the results of the two reference models that used all signals in the dataset, with an RMSE of 0.0145 mm and an R^2 score of 0.9967. Moreover, it is shown that the LSTM and BiGRU base learners also outperformed the reference models with RMSE of 0.0207 and 0.0149 mm, respectively. Thus, the efficiency of the data cleaning and augmentation process before training DL models was proven.

The performance results for the meta-learning model when trained with smcDC and AE signals showed a bigger margin of error with an RMSE of 0.0473 mm and an R^2 score of 0.9660. Nevertheless, the model required less inputs and the results are comparable to the reference models that use all signals. Finally, the results when using individual signals were underperforming. To achieve good results, the architecture of the models was expanded, adding two extra layers to the base learners. With smcDC, the model had an RMSE of 0.1699 mm and an R^2 score of 0.5715, and, with vib_spindle, the model had an RMSE of

Table 3. Hyperparameters and performance metrics of baseline models.

Model	Hyperparameters	R^2	RMSE	MAE
Decision tree	Default parameters	0.7225	0.1371	0.0534
SVM	C = 9.8143, ε = 0.0012, Kernel = RBF	0.8639	0.0961	0.0595
Random forest	Max. depth = 20, No. estimators = 437	0.8471	0.1018	0.0610
Gradient boosting	Learning rate = 0.0975, Max. depth = 13, No. estimators = 169	0.8570	0.0984	0.0623
XGBoost	Learning rate = 0.0098, Max. depth = 12, No. estimators = 577, Min. child weight = 4	**0.8952**	**0.0843**	**0.0478**
kNN	No. neighbours = 2, Weights = Distance	**0.9195**	**0.0739**	**0.0224**

Table 4. Performance results of the meta-learning model. Best performing models are highlighted in bold.

	All signals			DC and AE signals			DC signal			Vibration signals		
Model	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
LSTM	0.9935	0.0207	0.006	0.9454	0.0601	0.0225	0.3606	0.2078	0.1435	0.4712	0.1910	0.1348
CNN	0.9611	0.0507	0.0308	0.8597	0.0963	0.0632	0.5114	0.1815	0.1119	0.5778	0.1707	0.1169
BiGRU	0.9966	0.0149	0.0042	0.9630	0.0494	0.0229	0.3650	0.2067	0.1433	0.7997	0.1176	0.0748
Meta-learning	**0.9967**	**0.0145**	**0.0055**	**0.9660**	**0.0473**	**0.0220**	0.5715	0.1699	0.1048	0.8072	0.1130	0.0714
CNN with spectral subtraction [5]								**0.088**				
LSTM with process information [6]		0.0456	0.0322									
Hybrid LSTM [7]										**0.9837**	**0.0364**	**0.0258**
TM3C-KT [8]		0.0424										

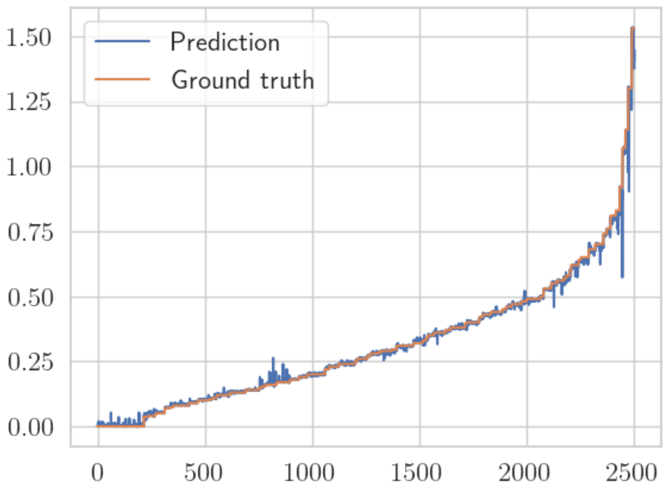

Fig. 5. Tool wear curve of the test data set, ordered by the ground truth VB value.

0.1130 mm and an R^2 score of 0.8072. Consequently, further research efforts should be given to improve the performance of the meta-learning model when using individual signals.

Fig. 5 presents a comparison of the VB curve for both the ground truth values, as well as the predicted values of the meta-learning model, when using all inputs. The data was ordered by ground truth value, as all cases and runs were augmented and shuffled stochastically during splitting. It may be observed that the model predicted values very close to the ground truth throughout the wear curve, proving the effectiveness and good performance of the approach when using sensor fusion.

5. Summary and conclusions

In this paper, a tool wear monitoring approach based on meta-learning using deep ensemble learning has been presented. The meta-learning approach is proposed for improving performance when predicting tool wear in machining. The approach uses deep ensemble learning to combine the outputs of multiple deep neural network models, i.e., LSTM, CNN, and BiGRU models, resulting in improved accuracy and robustness.

The meta-learning approach has been validated using the NASA Ames/UC Berkeley open access milling dataset, which was augmented and denoised. A combination of all the signals, smcDC and AE signals, and individual signals (smcDC and vib_table) were used for the validation tests. The best results were obtained when using all the signals, substantially outperforming state of the art DL-based reference models and proving the benefits of sensor fusion. Future work will involve investigating the ability of the meta-learning approach to detect tool wear in other machining datasets. Furthermore, data pre-processing and feature extraction techniques, as well as DL model hyperparameter tuning and architectural changes, will be studied to improve the performance of the approach when using individual signals as inputs.

Acknowledgements

This project has received funding from the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 814078 and by the Department of Education, Universities and Research of the Basque Government under the projects Ikerketa Taldeak (Grupo de Ingeniería de Software y Sistemas IT1519-22 and Grupo de investigación de Mecanizado de Alto Rendimiento IT1443-22).

References

[1] C.H. Lauro, L.C. Brandão, D. Baldo, R.A. Reis, J.P. Davim, Monitoring and processing signal applied in machining processes - A review, Measurement. 58 (2014) 73–86.
[2] R. Teti, D. Mourtzis, D.M. D’Addona, A. Caggiano, Process monitoring of machining, CIRP Annals. 71 (2022) 529–552.
[3] D.Y. Pimenov, A. Bustillo, S. Wojciechowski, V.S. Sharma, M.K. Gupta, M. Kuntoğlu, Artificial intelligence systems for tool condition monitoring in machining: analysis and critical review, J Intell Manuf. (2022) 1–43.
[4] A. Agogino, K. Goebel, Milling data set, (2007).
[5] F. Aghazadeh, A. Tahan, M. Thomas, Tool condition monitoring using spectral subtraction and convolutional neural networks in milling process, International Journal of Advanced Manufacturing Technology. 98 (2018) 3217–3227.
[6] W. Cai, W. Zhang, X. Hu, Y. Liu, A hybrid information model based on long short-term memory network for tool condition monitoring, J Intell Manuf. 31 (2020) 1497–1510.
[7] S. Kumar, T. Kolekar, K. Kotecha, S. Patil, A. Bongale, Performance evaluation for tool wear prediction based on Bi-directional, Encoder–Decoder and Hybrid Long Short-Term Memory models, International Journal of Quality and Reliability Management. 39 (2022) 1551–1576.
[8] S. Pillai, P. Vadakkepat, Deep learning for machine health prognostics using Kernel-based feature transformation, J Intell Manuf. 33 (2022) 1665–1680.
[9] P. Cawood, T. Van Zyl, Evaluating State-of-the-Art, Forecasting Ensembles and Meta-Learning Strategies for Model Fusion, Forecasting. 4 (2022) 732–751.
[10] S. Russel, P. Norvig, others, Artificial intelligence: a modern approach, Pearson Education Limited London, 2013.
[11] C. Bentéjac, A. Csörgő, G. Martínez-Muñoz, A comparative analysis of gradient boosting algorithms, Springer Netherlands, 2021.
[12] J. Wang, J. Xie, R. Zhao, L. Zhang, L. Duan, Multisensory fusion based virtual tool wear sensing for ubiquitous manufacturing, Robot Comput Integr Manuf. 45 (2017) 47–58.
[13] R. Vidal, Y. Ma, S.S. Sastry, Generalized Principal Component Analysis, Springer-Verlag New York, 2016.
[14] Z.-H. Zhou, Ensemble methods: foundations and algorithms, CRC press, 2012.
[15] L. Rokach, Pattern classification using ensemble methods, World Scientific, 2010.