Fusion of Feature Selection Methods for Improving Model Accuracy in the Milling Process Data Classification Problem

Maciej Kusy *Member, IEEE*
Faculty of Electrical and Computer Engineering
Rzeszow University of Technology
al. Powstancow Warszawy 12, 35-959 Rzeszow, Poland
mkusy@prz.edu.pl

Roman Zajdel
Faculty of Electrical and Computer Engineering
Rzeszow University of Technology
al. Powstancow Warszawy 12, 35-959 Rzeszow, Poland
rzajdel@prz.edu.pl

Jacek Kluska
Faculty of Electrical and Computer Engineering
Rzeszow University of Technology
al. Powstancow Warszawy 12, 35-959 Rzeszow, Poland
jacklu@prz.edu.pl

Tomasz Zabinski
Faculty of Electrical and Computer Engineering
Rzeszow University of Technology
al. Powstancow Warszawy 12, 35-959 Rzeszow, Poland
tomz@prz.edu.pl

***Abstract*—The current study addresses the problem of feature selection performed for the data set collected in the milling process. The data consists of 1709 records with 44 statistical parameters computed on the basis of the measured input signals from the accelerometer mounted on a lower bearing of the spindle of Haas VM-3 CNC machining centre and the acoustic emission sensor mounted in the machine cabin. A new feature selection approach is proposed which is based on the fusion of three filter methods: Pearson's linear correlation coefficient, ReliefF and single decision tree. By means of the introduced combined ranking set, the most significant features are stored and then employed to create the reduced data set. The validity of the proposed solution is tested by computational intelligence models in original and reduced data classification task. Based on the experimental study the efficacy of the approach is confirmed.**

***Index Terms*—feature selection, features' significance, combined ranking, neural networks, support vector machines, milling process**

I. Introduction

Feature selection (FS) involves creating a subset of attributes from the entire set of predictor variables. This results in a lower dimensionality of the input data space. When performing FS, no data transformation occurs – one retains original values of chosen attributes. For supervised classification tasks, FS approaches are split into filter and wrapper methods. In filter methods, FS process is isolated from the learning algorithm of a model. The relevant attributes are chosen based on assumed correlation between particular features and an output class. FS is usually conducted by means of FOCUS algorithm [1], fast correlation based filter approach [2], ReliefF [3] or a decision tree [4]. In the wrapper approach, which is not under investigation in the presented work, a classifier is involved in FS process since based on its performance most suitable feature subset is chosen.

In the literature, some attention has been paid to the feature selection in the tool condition monitoring, also in the milling processes. For example, in [5], decision tree (C4.5), scatter matrix, adaptive neuro-fuzzy inference system and a crosscorrelation method are utilized to find various features' subsets out of 25 available features from the data representing a wear of the face mill. Regular and entry cuts are investigated. The data are collected by acoustic emission sensors, accelerometers and motor current sensor to determine the state of the tool. A feed forward neural network is applied to evaluate five sensors-based FS schemes by comparing their classification rate and test errors. A significant improvement of the network's classification capabilities is observed after performing FS. The authors of [6] apply a modified Fisher's linear discriminant analysis for FS from cutting force signals acquired in the micro-milling process. The attributes are ranked according to their class–discriminant ability. The data with reduced number of features from 37 to 8 are used as the input for the hidden Markov model (HMM) in the classification task. It is shown that the proposed method improves HMM performance. The work [7] focuses on the fault diagnosis of the face milling tool during machining of a steel alloy. Vibration signals of the tool are acquired under healthy and different fault conditions. FS is performed with the use of the decision tree (J48) which selects 7 salient attributes out of all 30 histogram features extracted from the original signals. A K–star algorithm is then used as a classifier. After feature selection, high performance of the algorithm is achieved.

One must be aware of the fact that the use of a sole FS method in a considered attribute selection problem may prove to be inadequate for a given model. The resulting feature subset may turn out to be nonoptimal. Also, a selected method may be applicable to a certain task while the other one can be unsuitable. Thus, choosing a single best method for feature selection is problematic.

However, as shown in many FS related scientific papers, an in-depth investigation has been carried out to determine whether a fusion of state-of-the-art FS methods can enhance prediction capabilities of tested models. Rokach et al. [8]

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

present how an ensemble approach can be applied to improve FS performance. They show a general framework for creating the subsets of attributes which are further combined into a single final set. The combination is realized with the use of the scheme based on voting. In [9], an algorithm of merging various FS approaches is provided. It is based on the combinatorial fusion analysis model. A rank-score function and an associated graph are used to determine the diversity among applied FS methods. The work in [10] presents the algorithm which allows an ensemble of FS methods to reject detrimental attributes. It uses feature "rankers" which determine the list of features sorted based on their importance. The attributes are then collected into a single list with the use of a suitable aggregation function which provides a score for each feature based on the feature's placing in the original ranking list. Cateni et al. [11] introduce a novel combination of three filter FS algorithms: Fisher criterion, T–test and Kullback-Leibler divergence. Each method calculates attribute scores which are then appropriately combined to determine the mean value. Based on the mean and assumed threshold, reduced feature set is created. Finally, an exhaustive search is conducted to obtain a sub-optimal set of variables. In [12], a framework of methods for constructing ensembles of feature rankings is investigated. The methods take as input attribute rankings, generated by selected FS algorithms, and provide a feature relevance using various rank aggregation procedures. For experimental purposes, four different aggregation approaches are explored.

Based on the significant and justified contribution of fusion based feature selection methods in the field of attribute importance estimation, in this paper we propose new FS approach. It is based on Pearson's linear correlation coefficient, ReliefF and single decision tree algorithms. FS is applied to the data set represented by 44 parameters extracted from the signals acquired in the real milling process. The proposed approach, with the use of a combined ranking idea, collects the most significant attributes provided by each of the aforementioned methods. The obtained features are then fed as the input to three data classifiers: the multilayer perceptron (MLP), the probabilistic neural network (PNN) and the support vector machine (SVM). The accuracy of the classifiers computed on the data set with features obtained by the proposed approach as well as three independently utilized FS methods and original 44 attributes is compared. It is shown that the solution introduced in the current study contributes to the highest performance of the models.

The work presented in this paper is a part of a project aimed at developing efficient classification methods which can be used in real-time intelligent milling diagnostic system [13]. Milling is still a substantial technique used in industrial production, despite new developments in this field, i.e. 3D printing. Practical implementation of Industry 4.0 concept (i.e. automation and robotisation of manufacturing processes) requires application of intelligent diagnostic methods in technological processes supervision systems. In consequence, presented work has great practical significance.

This paper is organized as follows. Section II puts forward the architecture of the platform used for data acquisition, the way the milling experiments are conducted and the description of the extracted features. In Section III, the filter methods utilized for FS are described. In Section IV, the proposed FS method is introduced. The comparative analyses of the obtained results are presented in Section V. Finally, Section VI concludes current work.

II. Machining process and data representation

This section describes the data acquisition environment employed in the milling process. The features extracted from the measured signals used for the analyses are also presented.

A. Testbed

The testbed consists of a 3–axis vertical CNC machining center, a set of sensors and a data acquisition system. Milling experiments are performed on Haas VM-3 CNC machining center equipped with an inline direct-drive spindle. The set of sensors includes: 7 accelerometers (single-axis), 1 acoustic emission sensor and 1 force and torque sensor (thee-axis). In this study, the signals collected from two sensors are used, i.e. the accelerometer (ACC) mounted on a lower bearing of the spindle (sensitivity 100 mV/g, bandwidth 10 kHz) and the acoustic emission (AE) sensor mounted in the machine cabin (sensitivity 53 mV/Pa). The data acquisition process is performed by a platform for rapid prototyping of intelligent diagnostic systems developed by the authors [13]. The platform includes Beckhoff Industrial Computer C6920 (IPC) and the distributed input/output system based on EtherCAT protocol. To collect signals from ACC and AE sensors, the analog input modules EL3632 from Beckhoff are applied. The modules EL3632 are designed for devices which meet the Integrated Electronics Piezo-Electric standard. The oversampling factor of EL3632 module is set to 50. The software part of the data acquisition system consists of a real-time PLC task created in Structured Text language (norm IEC 61131-3) with the use of a factory automation programming environment TwinCAT 3 form Beckhoff as well as custom made Simulink projects and Matlab scripts. Data collection performed during milling experiments is done using Matlab/Simulink External Mode. The sampling interval of the IPC real-time data collection task and the duration of signal buffer (time series data) are equal to 2 ms and 640 ms, respectively. Taking into account the IPC real-time collection task interval and the EL3632 oversampling factor, the final sampling interval for the signals collected from the ACC and AE sensors takes the value of 40 μs (25 kHz sampling frequency). 16000 samples from each sensor are stored in each data buffer.

B. Milling experiments

In this study, the data from 11 experiments are analyzed. A single experiment includes the time series data collected from ACC and AE sensors for one complete circular milling trajectory performed at the edge of one Inconel 625 disc (diameter: 100 mm, thickness: 8 mm). Each machining experiment, lasting approximately 100 seconds, is performed

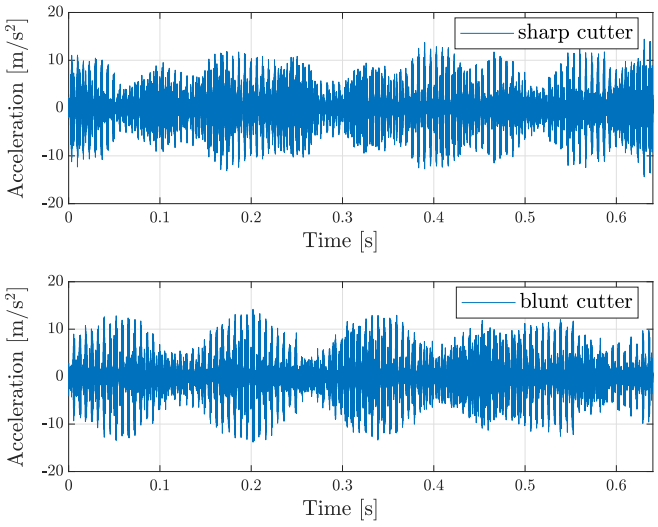

Fig. 1: The exemplary time plots for acceleration signal – sharp and blunt cutters.

with the use of the same one four-teeth milling cutter and the spindle speed 862 rpm (revolutions per minute) which corresponds to 14.36 Hz. Two types of milling cutters are used during the experiments, i.e. sharp and blunt. After each milling experiment, the roughness of the disk surface is measured; it is approximately 0.5 μm and 1 μm for sharp and blunt tools, respectively. In the current work, data buffers collected for machining done with different pieces of sharp tools (7 discs) are treated equally. The same approach is applied in case of different pieces of blunt tools (4 discs). Finally, after cleaning and pre-processing operations performed for all collected data, 1085 and 624 buffers are used for the analysis of the machining process done with sharp and blunt tools respectively. Exemplary plots for time series data collected for ACC sensor during machining with sharp and blunt tools are shown in Fig. 1.

C. Features Extracted From the Input Signals

The main purpose of feature extraction is to significantly reduce the dimension of raw data in time and maintain the relevant information in the extracted features. Many research works have investigated various feature extraction methods [14], [15], [16], [17]. In this paper, a set of features in time domain obtained from ACC and AE sensors are studied and defined. They are summarized in Table I. The amplitude values of both signals are expressed as $x_1, x_2, \ldots, x_n$, where $n = 16000$. Finally, 44 features are extracted from the signals collected by means of both sensors.

This work leaves out many problems, such as feature construction, embedded and hybrid feature selection, individual vs. subset feature evaluation, time complexity of the proposed procedure, as well as time-series analysis techniques. Even though the data are time-series, the authors try to make the method proposed in this article applicable to other data sets, not just time-series. Moreover, the results of this work allow us to assess which groups of features of the considered time-series in time or frequency domain are more or less important if the time-series analysis approach is chosen in the milling process data classification problem.

III. Feature Selection Approaches

This section outlines three filter methods used for feature subset selection among all 44 available attributes described in Section II and defined in Table I. At the end of this part of the work, the relevance of the features determined by considered filter methods is discussed.

A. Feature relevance based on Pearson's linear correlation coefficients

For the i-th feature, a Pearson's linear correlation coefficient (PLCC) is defined as the covariance of the variables divided by the product of their standard deviations [18]:

$$r_i = \frac{\sum_{l=1}^{L}(x_{li}-\bar{x}_i)(t_l-\bar{t})}{\sqrt{\sum_{l=1}^{L}(x_{li}-\bar{x}_i)^2}\sqrt{\sum_{l=1}^{L}(t_l-\bar{t})^2}}, \tag{1}$$

where $\bar{x}_i$ is the mean value of the i-th feature over all input data: $\bar{x}_i = L^{-1}\sum_{l=1}^{L} x_{li}$, and $\bar{t}$ is the mean target value: $\bar{t} = L^{-1}\sum_{l=1}^{L} t_l$. One can show that the values of r_i are independent of outputs coding, and $r_i = \pm 1$ if the vectors $[x_{1i},\ldots,x_{Li}]^T$ and $[t_1,\ldots,t_L]^T$ are linearly dependent, and 0 if they are linearly uncorrelated. Since the probability that two variables are correlated is established based on the complementary error function $P = \mathrm{erfc}\left(|r|\sqrt{I/2}\right)$ which is monotonically decreasing function of $|r|$, the feature relevance ranking can be obtained using the values of P or $|r|$.

B. ReliefF algorithm

ReliefF, proposed in [3], computes a weight vector whose elements provide the significance of particular data features. Formally, it finds K nearest neighbors for the record $\mathbf{x}_l$ among: (i) data of the same class: $[\mathbf{h}_1,\ldots,\mathbf{h}_k,\ldots,\mathbf{h}_K]$ – nearest hits; (ii) data from the remaining classes: $[\mathbf{m}_1^{(j)},\ldots,\mathbf{m}_k^{(j)},\ldots,\mathbf{m}_K^{(j)}]$ – nearest misses; $j = 1,\ldots,J$ and $j \neq c$ where c refers to the class of $\mathbf{x}_l$. The feature weights are determined as follows:

$$\begin{aligned} w_i^{\text{new}} &= w_i^{\text{old}} - \sum_{k=1}^{K}\frac{\Delta(x_{li},h_{ki})^2}{L\cdot K} \\ &+ \sum_{j=1,\, j\neq c}^{J}\left[\frac{P(j)}{1-P(c)}\sum_{k=1}^{K}\Delta\left(x_{li},m_{ki}^{(j)}\right)^2\right]/(L\cdot K), \end{aligned} \tag{2}$$

where $P(j)$ and $P(c)$ are the occurrence probabilities of class j and c, respectively and Δ calculates the difference between the values of the i-th feature for two records; $\Delta \in \{0,1\}$ for discrete features while for continues values $\Delta \in [0,1]$ [3]. The greater weight value, the higher significance of a feature. The choice of K influences the obtained relevance. In [19], $K = 10$ is recommended.

TABLE I: The list of features extracted from the input signals.

Index of the feature			
ACC	AE	Feature description	Expression
1	23	Maximum	$\max = \max_{i=1,\ldots,n} \{x_i\}$
2	24	Minimum	$\min = \min_{i=1,\ldots,n} \{x_i\}$
3	25	Peak to peak	$P = \max - \min$
4	26	Median	"middle" value in the sample
5	27	Maximum of the absolute value	$\max_a = \max_{i=1,\ldots,n} \{\lvert x_i \rvert\}$
6	28	Mean	$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$
7	29	Mean of the absolute value	$\mu_a = \frac{1}{n}\sum_{i=1}^{n} \lvert x_i \rvert$
8	30	Variance	$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2$
9	31	Root mean square	$RMS = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$
10	32	Standard deviation	σ
11	33	Energy	$E = \sum_{i=1}^{n} x_i^2$
12	34	Energy of the centered signal	$E_c = \sum_{i=1}^{n} (x_i - \mu)^2$
13	35	Kurtosis	$K = \frac{m_4}{\sigma^4}$
14	36	Skewness	$S = \frac{m_3}{\sigma^3}$
15	37		
16	38		
17	39	k-th order moment	$m_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^k$
18	40	for $k = 5, \ldots, 10$	
19	41		
20	42		
21	43	Shannon entropy	$I = -\sum_{i=1}^{n} x_i^2 \log_2 x_i^2$
22	44	Signal rate	$S = \frac{P}{\mu}$

C. Single decision tree

Decision trees make splits that maximize the decrease in impurity. By calculating the mean decrease in impurity for each feature across all trees we can know that feature importance. Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature.

D. Generated importance of the features

Table II shows the indices of 44 features ranked by PLCC, ReliefF and SDT in terms of their decreasing importance. As delineated, there are only 19 features specified by SDT since the grown tree selected only 19 features as nodes rejecting remaining 25 attributes. The first row of the table indicates that all three FS methods provide different feature as the most significant, i.e. m_5 for the ACC and μ_a for both the ACC and AE sensors.

IV. Proposed Feature Selection Method

The main purpose of this study is to propose an approach that allows us to determine the most relevant ordered subset of features, called a combined ranking score set. However, the total number of ordered subsets of the set $\{1,\ldots,I\}$ is $|\mathscr{N}(I,k)| = \sum_{j=0}^{k} I!/(I-j)!$, where $I = 44$ in our case, and k is cardinality of $\mathscr{N}$. For example, $|\mathscr{N}(44,3)| = 81401$ and $|\mathscr{N}(44,15)| \approx 3.1 \cdot 10^{23}$. Thus, an exhaustive search method for an optimal combined ranking score set is unfeasible. In order to solve this issue, three different filter methods are applied. However, as presented in Section III, each method provides different outcome which is reflected in diversified features importance order (see Table II). One can state the problem whether it is possible to find a subset of features that is an effect of applying various filter methods at the same time. Therefore, in this section we are interested in a fusion of three different FS methods. As a result we find a suboptimal subset of features which selects the most relevant ones, utilizing the results of PLCC, ReliefF and SDT algorithms. This solution is based on simple combined ranking score criterion which takes into account the particular position of the feature in the

TABLE II: The sets of features' indices sorted according to the importance provided by PLCC ($\mathscr{P}$), ReliefF ($\mathscr{R}$) and SDT ($\mathscr{T}$). The descending order is preserved meaning, the first feature is the most significant, as denoted by rank column N.

N	$\mathscr{P}$	$\mathscr{R}$	$\mathscr{T}$	N	$\mathscr{P}$	$\mathscr{R}$
1	15	7	29	20	20	38
2	14	14	7	21	3	36
3	17	29	8	22	5	15
4	19	35	15	23	33	39
5	29	28	13	24	30	43
6	10	32	17	25	34	21
7	9	31	32	26	36	37
8	8	25	23	27	22	22
9	12	24	5	28	37	11
10	11	27	31	29	39	12
11	7	13	6	30	38	8
12	31	23	28	31	6	9
13	32	26	25	32	28	10
14	16	34	20	33	41	16
15	1	30	33	34	44	44
16	27	33	1	35	40	17
17	18	42	18	36	42	3
18	23	40	2	37	2	18
19	25	41	24	38	26	19
				39	43	2
				40	13	5
				41	24	1
				42	21	20
				43	4	6
				44	35	4

significance order. Two following definitions are required to select a feature as important.

Definition 1: Let $\mathscr{P}$, $\mathscr{R}$ and $\mathscr{T}$ denote the sets of features' indices ordered according to the criterion of significance determined by PLCC, ReliefF and SDT methods, respectively. Let $\mathscr{P}_i$, $\mathscr{R}_i$ and $\mathscr{T}_i$ denote the subsets of the first i elements of $\mathscr{P}$, $\mathscr{R}$ and $\mathscr{T}$, respectively. Since not all the features may be included as tree nodes by SDT, it is assumed that $|\mathscr{P}| = |\mathscr{R}| > |\mathscr{T}|$, where $|\cdot|$ is the set's cardinality. The set of common features' indices is defined as follows:

$$\mathscr{C}_i = [(\mathscr{P}_i \cup \mathscr{R}_i) \cap \mathscr{T}] \cup \mathscr{T}_i, \tag{3}$$

where $i = 1, \ldots, |\mathscr{T}|$.

Definition 2: Given the sets $\mathscr{P}$, $\mathscr{R}$, $\mathscr{T}$ and the set of common features' indices $\mathscr{C}_i$. Let $\mathscr{C}_i^j$ denote some feature that is the j-th element of $\mathscr{C}_i$. Let $\mathscr{X}\left\{\mathscr{C}_i^j\right\}$ be some natural number which directly corresponds to the index of $\mathscr{C}_i^j$ in $\mathscr{X}$ where $\mathscr{X}$ stands for any of predefined sets of features' indices. A combined ranking score which is determined for the feature $\mathscr{C}_i^j$ selected in $\mathscr{P}_i$, $\mathscr{R}_i$ and $\mathscr{T}_i$ simultaneously is defined as follows:

$$R_{\mathscr{C}_i^j} = \sum_{s=1}^{3} \left(|\mathscr{T}| - \mathscr{X}_s\left\{\mathscr{C}_i^j\right\} + 1 \right) \tag{4}$$

for $\mathscr{X}_s\left\{\mathscr{C}_i^j\right\} \leqslant |\mathscr{T}|$. In (4), $\mathscr{X}_{s=1} = \mathscr{P}$, $\mathscr{X}_{s=2} = \mathscr{R}$, $\mathscr{X}_{s=3} = \mathscr{T}$ and $j = 1, \ldots, |\mathscr{C}_i|$. For any $\mathscr{X}_s\left\{\mathscr{C}_i^j\right\} > |\mathscr{T}|$, the s-th summand is not considered in computing $R_{\mathscr{C}_i^j}$. Adding 1 ensures assignment of the score from the set $\{1, \ldots, |\mathscr{T}|\}$ $\forall s$.

As the working example, it is convenient to use the sets $\mathscr{P}$, $\mathscr{R}$ and $\mathscr{T}$ presented in Table II. Let us consider $i = 3$ first elements of these sets; this means that 3 features are treated as important as the effect of applying PLCC, ReliefF and SDT. Then: $\mathscr{P}_3 = \{15, 14, 17\}$, $\mathscr{R}_3 = \{7, 14, 29\}$ and $\mathscr{T}_3 = \{29, 7, 8\}$. Since 14 does not occur in $\mathscr{T}$, $(\mathscr{P}_3 \cup \mathscr{R}_3) \cap \mathscr{T} = \{15, 7, 17, 29\}$. The set of common features' indices is therefore equal to: $\mathscr{C}_3 = \{15, 7, 17, 29, 8\}$. If one regards $j = 1$, $\mathscr{C}_3^1 = 15$ and therefore: $\mathscr{P}_3\{15\} = 1$, $\mathscr{R}_3\{15\} = 0$ and $\mathscr{T}_3\{15\} = 4$. The combined ranking score for the feature 15 chosen in $\mathscr{P}_3$, $\mathscr{R}_3$ and $\mathscr{T}_3$ simultaneously takes the value: $R_{15} = 19 - 1 + 1 + 19 - 4 + 1 = 35$. Similarly, $R_7 = 46$, $R_{17} = 31$, $R_{29} = 51$ and $R_8 = 29$. Based on the values of R, one obtains the following combined ranking set $\mathscr{R}_{\mathscr{C}_3} = \{29, 7, 15, 17, 8\}$. $\mathscr{R}_{\mathscr{C}_3}$ determines the indices of the features from most to least significant in $\mathscr{C}_3$.

Table III (the upper part) presents the combined ranking sets $\mathscr{R}_{\mathscr{C}_i}$. Note that the values of i are not successively increased by 1. This is explained as follows; if we choose, for example, $i = 1$ first elements of $\mathscr{P}$, $\mathscr{R}$ and $\mathscr{T}$ then $\mathscr{R}_{\mathscr{C}_1} = \{29, 7, 15\}$. However, taking $i = 2$ implies $\mathscr{R}_{\mathscr{C}_2} = \mathscr{R}_{\mathscr{C}_1} \cup \{14, 14, 7\}$. Since the feature 14 does not occur in $\mathscr{T}$ and the 7-th attribute is already included in $\mathscr{R}_{\mathscr{C}_1}$, $\mathscr{R}_{\mathscr{C}_2} = \mathscr{R}_{\mathscr{C}_1}$. Therefore, $i = 2$ is excluded from the table. Thus, from $i = 1, \ldots, |\mathscr{T}|$ of all possible values, we solely obtain $\{1, 3, 5, 6, 7, 8, 9, 11, 14, 15, 17, 18\}$ as the desired subset of indices i.

V. Simulation Results

In this section, we present the performance of MLP [20], PNN [21] and SVM [22] in the classification of considered input records with the FS applied to reduce data dimensionality. Each of these models operates on the training data which are regarded in terms of input–output pairs $(\mathbf{x}_l, t_l)$, where $\mathbf{x}_l = [x_{l1}, \ldots, x_{lI}]^T$ is the feature vector and t_l is its associated target. The data cardinality, the number of features and the number of given classes are equal to $L = 1709$, $I = 44$ and $J = 2$, respectively. For the purpose of the analysis, the following two tasks are regarded, i.e.: (i) the features are selected based on three filter methods and (ii) the features are chosen by means of the proposed approach. Additionally, the models are tested in the classification of the original data set consisting of 44 attributes. The performance is evaluated with the use of the classification accuracy:

$$Acc = \frac{1}{L} \sum_{l=1}^{L} \delta\left[y(\mathbf{x}_l) = t_l\right], \tag{5}$$

where $y(\mathbf{x}_l)$ is the classifier's output calculated for $\mathbf{x}_l$. In (5), $\delta[\cdot] = 1$ when $y(\mathbf{x}_l) = t_l$ and 0, otherwise. Acc is determined using a 10–fold cross validation procedure.

A. Parameter settings

In the experiments, both FS methods and data classifiers need to be adjusted to provide the highest possible accuracy.

TABLE III: The upper part: the combined ranking sets for the indices of particular common features stored in $\mathscr{C}_i$. The lower part: the accuracy values (in %) achieved by MLP, PNN and SVM for the attributes included in $\mathscr{R}_{\mathscr{C}_i}$.

	N	$\mathscr{R}_{\mathscr{C}_1}$	$\mathscr{R}_{\mathscr{C}_3}$	$\mathscr{R}_{\mathscr{C}_5}$	$\mathscr{R}_{\mathscr{C}_6}$	$\mathscr{R}_{\mathscr{C}_7}$	$\mathscr{R}_{\mathscr{C}_8}$	$\mathscr{R}_{\mathscr{C}_9}$	$\mathscr{R}_{\mathscr{C}_{11}}$	$\mathscr{R}_{\mathscr{C}_{14}}$	$\mathscr{R}_{\mathscr{C}_{15}}$	$\mathscr{R}_{\mathscr{C}_{17}}$	$\mathscr{R}_{\mathscr{C}_{18}}$
Indices	1	29	29	29	29	29	29	29	29	29	29	29	29
	2	7	7	7	7	7	7	7	7	7	7	7	7
	3	15	15	15	15	15	15	15	15	15	15	15	15
	4		17	17	32	32	32	32	32	32	32	32	32
	5		8	8	17	17	17	17	17	17	17	17	17
	6			13	8	31	31	31	31	31	31	31	31
	7			28	13	8	8	8	8	8	8	8	8
	8				28	13	13	13	13	13	13	13	13
	9					28	28	28	28	28	28	28	28
	10						23	23	23	23	23	23	23
	11						25	25	25	25	25	25	25
	12							24	24	24	24	24	24
	13							5	5	5	5	5	5
	14								6	6	1	1	1
	15									20	6	6	6
	16										33	33	33
	17										20	18	18
	18											20	20
	19												2
Accuracy	Classifier												
	MLP	83.39 ±1.53	84.55 ±1.31	**87.67 ±1.28**	87.14 ±1.62	87.03 ±1.67	85.66 ±2.44	84.97 ±1.82	84.87 ±2.17	84.54 ±2.40	84.32 ±2.17	84.62 ±2.13	84.21 ±2.08
	PNN	86.01 ±0.18	86.43 ±0.15	**88.42 ±0.23**	87.99 ±0.36	88.02 ±0.35	87.62 ±0.18	87.91 ±0.60	87.62 ±0.35	87.88 ±0.22	87.55 ±0.14	87.81 ±0.24	87.36 ±0.33
	SVM	85.37 ±0.32	86.52 ±0.16	88.29 ±0.22	88.52 ±0.29	**88.77 ±0.14**	88.60 ±0.10	87.99 ±0.40	86.60 ±0.43	86.14 ±0.33	85.20 ±0.34	85.78 ±0.45	85.38 ±0.29

For this purpose, the appropriate grid search is performed and the following parameter settings are applied:

1) For FS approaches:
   - ReliefF: number of the nearest neighbors: $K = \{6,8,10,12\}$.
   - SDT: splitting algorithm: entropy; minimum rows in a node: 10; maximum depth: 10.
2) For classification models:
   - MLP: single-hidden-layer network with the number of neurons $H_1 \in \{2,4,6,\ldots,100\}$ and two-hidden-layer network where $H_1 > H_2$; training algorithm: conjugate gradients, Levenberg–Marquardt method;
   - PNN: smoothing parameter in the form of a matrix with elements referring to each class and each feature adjusted by means of conjugate gradient procedure.
   - SVM: kernel function: (i) polynomial with $d = \{2,3,4,5\}$, (ii) Gaussian with $\sigma \in [0.05,50]$; $C = \{10^0,10^1,10^2,10^3,10^4,10^5\}$; quadratic programming problem solved by sequential minimal optimization.

The simulations are run in Matlab and DTREG software.

B. Application of filter methods

The purpose of the current analysis is to show how the FS conducted by means of PLCC, ReliefF and SDT methods affects the classification accuracy of MLP, PNN and SVM. Given the significance of the attributes, which are ordered from the most to the least relevant, the considered data classifiers are firstly applied in the classification task of all input examples with only one, most important feature. Thereafter, a single less meaningful feature is added to the input space and data classification is performed. The procedure is repeated until all attributes (44 for PLCC, ReliefF and 19 for SDT) are presented to the classifiers. Fig. 2 (a), (b) and (c) illustrate the influence of the features' significance provided by filter methods on the accuracy of MLP, PNN and SVM, respectively. The following observations are worth stressing:

1) The accuracy changes follow similar pattern for each model: low initial values, then abrupt increase and settling within small changes for a greater number of features (N).
2) For each model, some analogy resulting from the application of FS takes place:
   - PLCC: four low initial values of Acc at $\approx 65\%$; sudden increase of Acc to $\approx 82\%$; 82% accuracy level for $N = \{6,7,8,9,10\}$; further gain in the accuracy up to $\approx 87\%$ for $N > 12$.
   - ReliefF: two low early accuracy values; sudden increase of Acc to $\approx 85\%$ followed by its constant level for $N = \{3,\ldots,25\}$; for $N > 25$, $\approx 5\%$ growth of the accuracy for MLP and SVM;
   - SDT: single low $Acc \approx 75\%$ for $N = 1$ followed by sudden increase up to $\approx 85\%$; subsequent accuracy gain to $\approx 87\%$; minor decrease of Acc for $N > 10$.

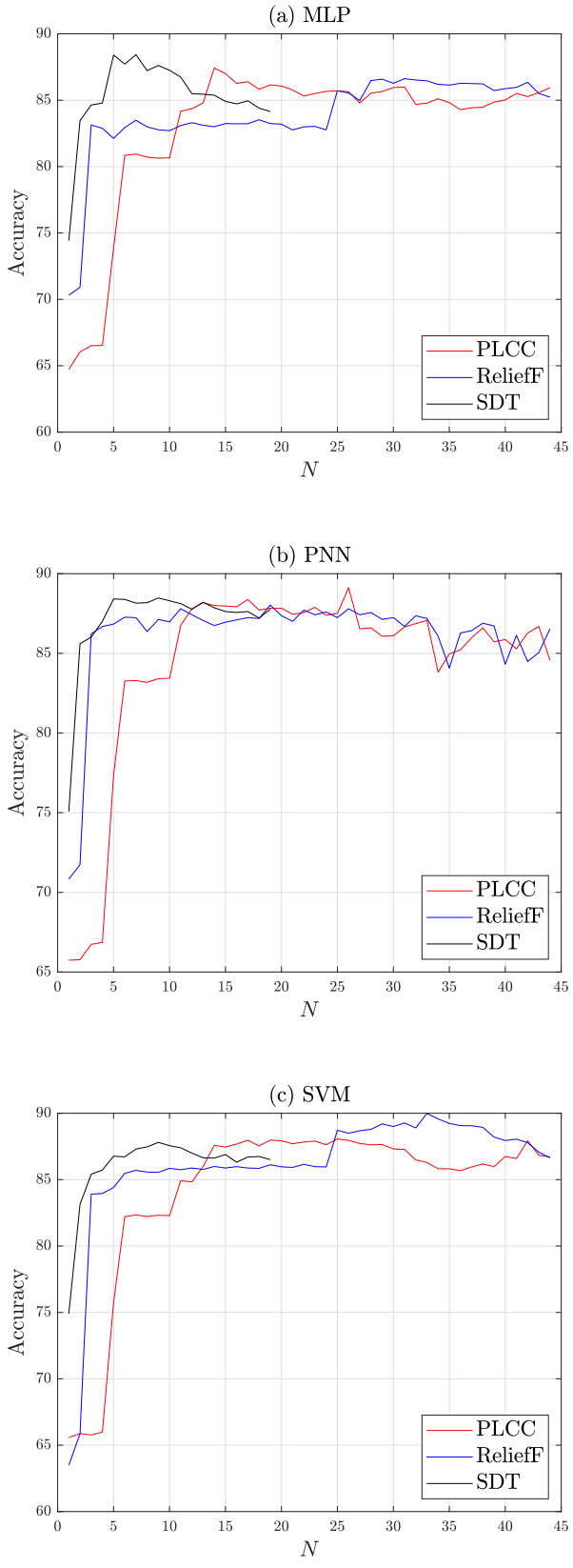

Fig. 2: Averaged accuracy for MLP, PNN and SVM. The axes labeled N represents the number of features which constitute the input for all classifiers.

Table IV presents the highest accuracy results attained by MLP, PNN and SVM for the input records having features reduced by the PLCC, ReliefF and SDT methods (columns 2, 3 and 4, respectively). Next to the *Acc* values, we indicate the number of features for which a particular outcome is obtained. As shown, for each filter method, the accuracies of data classifiers achieve higher rates than those for original 44 input space (last column in the table). Also note that this improvement takes place when $N < 44$.

C. Features' selection based on combined ranking set

The data sets with features' subsets provided by the combined ranking method (Table III) are used to train considered data classifiers.

The remarks presented below stress the validity and efficiency of the proposed FS approach. From the results shown in Table IV we can conclude as follows:

1) There exist such a data set with the features represented by $\mathscr{R}_{\mathscr{C}_i}$ for which the accuracy values of MLP, PNN and SVM are higher than *Acc* for the original data set.
2) The number of features provided by the proposed method is significantly lower than N obtained by the PLCC and ReliefF methods for all classifiers. However, for MLP, $N = |\mathscr{R}_{\mathscr{C}_5}|$ is equal to 7 which is the number of the attributes determined by SDT, while for SVM model $N = |\mathscr{R}_{\mathscr{C}_7}| = 9$. Admittedly, the proposed methods returns $N = 7$ for PNN while $N = 6$ for SDT, but the PNN's accuracy value is higher than the one of the SDT model.
3) For each data classifier, the accuracy achieved on the data set reduced by the proposed method is higher than *Acc* obtained after applying two out of three filter methods.
4) For all classification models, the highest accuracy is determined by a different base FS method. There is no single FS method that provides the highest *Acc* for at least two models.

If we have a closer look to the bottom three rows in Table III, one observes that achieving maximum value of *Acc* for MLP and PNN at $N = 7$ is influenced by the features indexed 13 and 28 which occurred in $\mathscr{R}_{\mathscr{C}_5}$. Similarly, those two features contribute to the improvement of the SVM's accuracy. However, the highest performance of SVM is attained for $\mathscr{R}_{\mathscr{C}_7}$, i.e. when the features 32 and 31 are also included in the input space. It is also worth noting that the accuracies of MLP and SVM steadily decrease their values after the highest *Acc* is determined ($\mathscr{R}_{\mathscr{C}_5}$ and $\mathscr{R}_{\mathscr{C}_7}$, respectively) to reach over 3% lower rates for $\mathscr{R}_{\mathscr{C}_{18}}$. In the case of a PNN, such a decrease is equal to over 1%.

VI. Conclusions

In this article, the method for the attributes' selection was proposed. The utilized data set took the form of 44 parameters extracted from the input signals acquired within the milling process. By merging three filter methods: PLCC, ReliefF and SDT the proposed method utilized the combined ranking score to provide the set of the most relevant features. To verify the validity of the introduced approach, MLP, PNN and SVM were

TABLE IV: First three columns: the highest *Acc* for MLP, PNN and SVM obtained on the data set with the number of attributes reduced to *N* according to three FS procedures; fourth column: the highest *Acc* for the combined ranking set $\mathscr{R}_{\mathscr{C}_i}$; final column: *Acc* for the entire data set. The outcomes (in %) are averaged over 10 simulation runs; standard deviations are included.

	PLCC		ReliefF		SDT		Proposed method		All features	
	Acc	*N*	*Acc*	*N*	*Acc*	*N*	*Acc*	*N*	*Acc*	*N*
MLP	87.43 ±1.60	14	86.62 ±2.09	31	88.43 ±1.52	7	87.67 ±1.28	7	85.94 ±1.87	44
PNN	88.75 ±0.24	26	88.18 ±0.24	26	88.34 ±0.18	6	88.42 ±0.23	7	84.97 ±0.73	44
SVM	88.06 ±0.44	25	89.98 ±0.23	33	87.81 ±0.16	9	88.77 ±0.14	9	86.81 ±0.01	44

applied to the classification tasks of the studied data set with features provided by the proposed method, PLCC, ReliefF, and SDT and also all training attributes.

The propounded heuristic fusion of feature selection methods can be perceived as universal since, due to the fact of the aggregation of the results obtained by the base filter algorithms, it is unnecessary to conduct the comparative experiments aiming at the choice of the FS method predisposed to a particular classification problem. The solution provided in this paper is concerned with three state-of-the-art approaches; however, it can be applied to other FS methods or a greater number of theirs.

The future work will focus on the task of weighting the features selected as significant by individual FS methods. Some criterion will be assumed for this purpose, e.g classification correctness of the model attained for a data with the subset of features selected by the base methods.

ACKNOWLEDGMENT

This work is financed by Polish Ministry of Science and Higher Education under the program "Regional Initiative of Excellence" in 2019–2022. Project number 027/RID/2018/19, funding amount 11 999 900 PLN.

REFERENCES

[1] H. Almuallim and T. Dietterich, "Learning with many irrelevant features," in *The Ninth National Conference on Artificial Intelligence*, 1991, pp. 547–552.

[2] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *The 20-th International Conference on Machine Learning*, 2003, pp. 547–552.

[3] I. Kononenko, "Estimating attributes: analysis and extensions of Relief," in *European Conference on Machine Learning*. Springer, 1994, pp. 171–182.

[4] C. Cardie, "Using decision trees to improve case-based learning," in *The 10-th International Conference on Machine Learning*, 1993, pp. 25–32.

[5] K. Goebel and W. Yan, "Feature selection for tool wear diagnosis using soft computing techniques," in *The ASME International Mechanical Engineering Congress and Exhibition*, 2000, pp. 5–10.

[6] K. Zhu, G. Hong, and Y. Wong, "A comparative study of feature selection for hidden markov model-based micro-milling tool wear monitoring," *Machining Science and Technology*, vol. 12, no. 3, pp. 348–369, 2008.

[7] C. Madhusudana, H. Kumar, and S. Narendranath, "Condition monitoring of face milling tool using k-star algorithm and histogram features of vibration signal," *Engineering science and technology, an international journal*, vol. 19, no. 3, pp. 1543–1551, 2016.

[8] L. Rokach, B. Chizi, and O. Maimon, "Feature selection by combining multiple methods," in *Advances in Web Intelligence and Data Mining*. Springer, 2006, pp. 295–304.

[9] Y. Li, D. F. Hsu, and S. M. Chung, "Combination of multiple feature selection methods for text categorization by using combinatorial fusion analysis and rank-score characteristic," *International Journal on Artificial Intelligence Tools*, vol. 22, no. 02, p. 1350001, 2013.

[10] B. Pes, "Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains," *Neural Computing and Applications*, pp. 1–23, 2019.

[11] S. Cateni, V. Colla, and M. Vannucci, "A hybrid feature selection method for classification purposes," in *2014 European Modelling Symposium*. IEEE, 2014, pp. 39–44.

[12] R. C. Prati, "Combining feature ranking algorithms through rank aggregation," in *The 2012 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2012, pp. 1–8.

[13] T. Zabinski, T. Maczka, and J. Kluska, "Industrial platform for rapid prototyping of intelligent diagnostic systems," in *Trends in Advanced Intelligent Control, Optimization and Automation*, W. Mitkowski, J. Kacprzyk, K. Oprzedkiewicz, and P. Skruch, Eds. Springer International Publishing, 2017, pp. 712–721.

[14] A. G. Rehorn, J. Jiang, and P. E. Orban, "State-of-the-art methods and results in tool condition monitoring: a review," *The International Journal of Advanced Manufacturing Technology*, vol. 26, no. 7-8, pp. 693–710, 2005.

[15] M. Elbestawi, J. Marks, and T. Papazafiriou, "Process monitoring in milling by pattern recognition," *Mechanical Systems and Signal Processing*, vol. 3, no. 3, pp. 305–315, 1989.

[16] R. Silva, R. Reuben, K. Baker, and S. Wilcox, "Tool wear monitoring of turning operations by neural network and expert system classification of a feature set generated from multiple sensors," *Mechanical Systems and Signal Processing*, vol. 12, no. 2, pp. 319–332, 1998.

[17] D. Brezak, T. Udiljak, D. Majetic, B. Novakovic, and J. Kasac, "Tool wear monitoring using radial basis function neural network," in *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)*, vol. 3. IEEE, 2004, pp. 1859–1862.

[18] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Pearson Correlation Coefficient*. Springer Berlin Heidelberg, 2009, pp. 1–4.

[19] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, no. 1-2, pp. 23–69, 2003.

[20] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*, D. E. Rumelhart and J. L. McClelland, Eds. Cambridge, MA, USA: MIT Press, 1986, pp. 318–362.

[21] D. F. Specht, "Probabilistic neural networks," *Neural Networks*, vol. 3, no. 1, pp. 109–118, 1990.

[22] V. N. Vapnik, *The nature of statistical learning theory*. Springer-Verlag, 1995.