Reliable information fusion methods for condition monitoring

Von der Fakultät für Ingenieurwissenschaften,
Abteilung Maschinenbau und Verfahrenstechnik der
Universität Duisburg-Essen
zur Erlangung des akademischen Grades

einer
Doktorin der Ingenieurwissenschaften
Dr.-Ing.

genehmigte Dissertation

von

Sandra Rothe
aus
Velbert, Deutschland

Gutachter:
Univ.-Prof. Dr.-Ing. Dirk Söffker
Univ.-Prof. Dr.-Ing. Claus-Peter Fritzen

Tag der mündlichen Prüfung: 13. Juni 2019

Acknowledgment

First, I would like to express my sincere gratitude to my supervisor Univ.-Prof. Dr.-Ing. Dirk Söffker, for offering me an opportunity to carry out my research as well as to work as scientific co-worker at the Chair of Dynamics and Control (SRS) at Universtity of Duisburg-Essen. I am very grateful for his guidance, advices, and help.

I would like to thank Univ.-Prof. Dr.-Ing. Claus-Peter Fritzen for his effort in examining my thesis.

I am grateful to my colleagues at the Chair of Dynamics and Control for their helpful support. Special thanks to my former colleague and friend Sandra Schönhoff for the good cooperation during our time at the Chair of Dynamics and Control and your help.

Special thanks to my family and especially to my fiancé Sebastian Viehöfer, who always believed in me. Without you, I could not have finished my thesis.

Duisburg, June 2019 Sandra Rothe

Kurzfassung

In vielen Anwendungsbereichen müssen zuverlässige Entscheidungen getroffen werden, um die Sicherheit zu erhöhen, die Funktionalität zu gewährleisten oder Kosten zu senken. Insbesondere im Bereich der Zustandsüberwachung ist die Bewertung von Situationen, Bedingungen oder Zuständen die Hauptaufgabe des Überwachungssystems. Zahlreiche Klassifikationsalgorithmen wurden entwickelt, um die Klassifikation von Situationen, Bedingungen oder Zuständen zu unterstützen. Die Leistung einzelner Klassifikationsalgorithmen variiert unter anderem mit den verwendeten Datensätzen, den verfügbaren Trainingsdaten für überwachte Klassifikatoren und der Hyperparameterabstimmung. Das Problem, nur einen Klassifikator zu verwenden, der speziell für bestimmte Situationen trainiert wurde, aber für nicht trainierte, unbekannte Situationen nicht geeignet ist, wird durch die Verwendung von Fusionsmethoden überwunden. Die Zuordnungen einzelner Basisklassifikatoren werden kombiniert, um die Gesamtzuverlässigkeit zu verbessern. Ähnlich wie bei den Klassifikationsalgorithmen führt Fusion nicht für jeden Datensatz, jede Klassifikatorauswahl oder jede Fusionsmethode zu zufriedenstellenden Ergebnissen.

Die Analyse von Eigenschaften, die das Fusionsergebnis beeinflussen, ist das Hauptthema dieser Arbeit. Dies umfasst zum einen die Analyse mit unterschiedlichen Datensätzen, die verschiedene Eigenschaften aufweisen (wie z. B. Anzahl der Klassen, etc.) und aus verschiedenen Anwendungsbereichen stammen, unter anderem aus der Zustandsüberwachung. Diese Daten werden unter Verwendung verschiedener Standardklassifikatoren mit Standardparametern klassifiziert, um die Analyse der Klassifikatoren an sich von den Untersuchungen auszuschließen. Verschiedene Ensembleauswahlstrategien werden betrachtet, um den Einfluss der Klassifikatoreigenschaften der Ensemblemitglieder auf die Gesamtfusionsperformanz zu untersuchen. Die Leistung der einzelnen Basisklassifikatoren sowie das für die Fusion verwendete Leistungsmaß werden variiert, um den Einfluss auf das Gesamtergebnis zu zeigen. Die Eigenschaften der Fusionsmethoden sowie die Datenmerkmale werden in Bezug auf ihren Einfluss auf die Genauigkeit der Ergebnisse untersucht.

Die Ergebnisse zeigen, dass Fusion die Gesamtperformanz verbessern kann, jedoch nur unter bestimmten Umständen. In dieser Arbeit wird der Zusammenhang zwischen verschiedenen Ensembleauswahlstrategien, der optimalen Anzahl und Eigenschaften der augewählten Klassifikatoren, der Dateneigenschaften sowie den Eigenschaften der Fusionmethoden und der Gesamtzuverlässigkeit untersucht. Ziel ist, die Bedingungen unter denen die Fusionsmethoden zu einer Verbesserung der Gesamtzuverlässigkeit führen bzw. unter denen es keinen Sinn ergibt, Fusion anzuwenden, zu bestimmen.

Abstract

Reliable decisions have to be made in several application fields to enhance safety, ensure functionality, or save costs. Especially in the field of condition monitoring, the evaluation of situations, conditions, or states is the main task of the monitoring system. Several classification approaches were developed to assist classifying situations, conditions, or states. The performance of individual classification approaches vary among others with the applied data sets, available training data for supervised classification approaches, and hyperparameter tuning. To overcome the problem of using only one classifier especially trained for specific situations, but performing not well for not trained unknown situations, fusion methods combine the assignments of individual base classifiers to improve the overall reliability. Similar to the classification approaches, fusion is not working well for every data set, classifier ensemble, or fusion method.

The analysis of characteristics influencing the fusion performance is the main topic of this thesis. This includes the analysis using different data with several characteristics (e.g number of classes, etc.) and originated from different application fields, also from condition monitoring approaches. These data are classified using different standard classification approaches with default parameters to exclude the analysis of classification approaches from the considerations. Several ensemble selection strategies are considered to examine the impact of the classifier properties of the ensemble members on the overall fusion performance. The performance of the individual base classifiers as well as the performance measure used for fusion are altered to show the influence on the overall result. Fusion method characteristics as well as data characteristics are examined related to their amount of influence on the performance.

The results show, that fusion can improve the overall performance, but only for specific circumstances. In this thesis the relationship between different ensemble selection strategies, the optimal number and properties of the selected classifiers, the data characteristics as well as the properties of the fusion methods and the overall reliability is investigated. The aim is to determine the conditions under which the fusion methods lead to an improvement in the overall reliability or under which it makes no sense to apply fusion.

Contents

List of Figures

List of Tables

Nomenclature

Symbols

a	Flaw size
$\hat{a}$	Flaw size response
$a_{90/95}$	Maximum flaw size that could be missed with 90 % POD at 95 % confidence level
b	Binary characteristic function
bc	Borda count
BKS_l	Unit l of the Behavior Knowledge Space
c_i	Class i with $i = 1, \ldots, n_C$
c_l	Best representative class for BKS_l
C^k	Confusion matrix for classifier e_k
C^k_{ij}	Number of samples, where classifier e_k assigned class j and actual class is i.
ΔAcc	Accuracy gain
e	Fuzzy integral
e_k	Classifier k with $k = 1, \ldots, n_K$
f	Maximum likelihood function
g	Fuzzy density
$H(C)$	Entropy of classes
I	Fisher's information matrix
k	Number of folds using k-fold cross-validation
L	Logit
m	Basic assignments
m_S	Number of samples of one fold
μ	Mean of POD curve
n_C	Number of classes
n_K	Number of classifiers
n_l	Number of samples corresponding to unit BKS_l
$n_l(c_i)$	Number of samples belonging to class c_i corresponding to unit BKS_l
n_S	Number of samples
N	Number of classes receiving v
$N_{(Y=1)}$	Number of true classes at v
ω	Final degree of support
p^k_i	Precision of classifier e_k for class i
P^k	Probability matrix for classifier e_k
P^k_{ij}	Probability that the samples belongs to class i, if classifier e_k assigned it to class j.
pm_i	Probability mass function for class i

q_k	Assigned class by classifier e_k
$\hat{q}$	Fused class label
r_k	Assigned ranking vector by classifier e_k
$r_{k,i}$	Assigned rank of class c_i by classifier e_k
$\hat{r}$	Fused ranking vector
$\hat{r}_i$	Fused rank of class c_i
s_k	Assigned measurement vector by classifier e_k
$s_{k,i}$	Assigned support value of class c_i by classifier e_k
$\hat{s}$	Fused measurement vector
$\hat{s}_i$	Fused support value
σ	Standard deviation of POD curve
T	Number of correctly classified samples
τ_y	Standard deviation of regression line
Θ	Frame of discernment (set of all classes)
v	Specific score vector
v_i	Assigned score for class c_i
$v_{k,i}$	Assigned score of class c_i by classifier e_k
w	Weighted characteristic function
x	Considered sample
y	Line of best fit (regression line)
$y_{(a=0.95)}$	95 % Wald confidence bound on y
Y	Response value
z	Standardized deviation of regression line
Z	Data set
Z_j	Fold j of data set Z

Abbreviations

ACC	Accelerometer
ALL	All available classifiers in the considered analysis
ANN	Artificial Neural Network
BC	Borda Count
BCR	Bayesian Combination Rule
BKS	Behavior Knowledge Space
CC	Cross-Correlation
CWT	Continuous Wavelet Transformation
DCE	Dynamic Classifier Ensemble
DSC	Dempster-Shafer Combination
DWT	Discrete Wavelet Transformation
EMD	Empirical Mode Decomposition
FAR	False Alarm Rate
FC	Fictional Classifier
FI	Fuzzy Integral
FN	False Negative
FP	False Positive
HR	Highest Rank
HRM	Hot Rolling Mill
KNN	K-Nearest Neighbor
Laser	Laser Sensor
LR	Logistic Regression
nB	The n best classifiers
nM	The n medium-good classifiers
nW	The n worst classifiers
NDT	Non-destructive Testing
NSGA-II	Non-Dominated Sorting Algorithm II
MV	Majority Voting
POD	Probability of Detection
RKHS	Reproducing Kernel Hilbert Space
ROC	Receiver Operating Characteristic
SCE	Static Classifier Ensemble
SG	Strain Gauge
SHM	Structural Health Monitoring
STFT	Short-Time Fourier Transform
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
WV	Weighted Voting
WVD	Wigner-Ville Distribution

1 Introduction

1.1 Motivation and problem statement

The overall system reliability of complex or safety critical systems is of increasing importance. For evaluating situations, conditions, or states, classification approaches are widely used in a lot of application fields, such as condition monitoring, image recognition, or object detection. The field of Structural Health Monitoring (SHM), i.e. the process of implementing a damage detection and characterization strategy for monitoring engineering structures [KN05, FW07], increasingly has become an essential aspect of industrial practice to ensure the quality of products, safe operations, improved maintenance, and to save costs. Many engineering structures are approaching or exceeding their initial design life, making SHM relevant [FDN01]. In SHM, monitoring is mainly applied online for large structures [FW07]. Non-destructive testing in contrast is usually applied offline after damage localization though it is used for in situ monitoring of structures like pressure vessels, rails, aircraft components among others. The implementation of decision support systems within complex and safety critical applications strongly relies on the dependable accuracy of decisions. In the last decades several condition monitoring techniques are implemented to detect changes, faults, and local defects. These methods can be grouped into four categories: signal-based [ASS14], model-based [Soe+16, IB97], data driven [IB97], and hybrid approaches [Poo+17]. Considering the application of methods for data classification, the assessment of situations or conditions should be improved.

Modern monitoring systems aim to process raw sensor data automatically using statistical learning theory (i.e. pattern recognition techniques) to obtain diagnostic statements. For further use of the information to automate the decision process regarding management, monitoring, or maintenance tasks, high detection rates and low false alarm rates are required. In the field of Non-destructive Testing (NDT), well-founded theory has been developed to assess the reliability of testing procedures. The Probability of Detection (POD), which is a probabilistic approach to assess the reliability of an NDT method [Kur+13], is frequently used. The POD curve of an NDT inspection technique is computed with respect to a fixed decision threshold using model (calibration) specimens under controlled laboratory conditions [Man+11]. Online diagnosis of machinery requires in service application. Here, damages evolve over time and disturbances are generally possible [CFM09]. Consequently, the sensor output is compared to a baseline signal for damage detection, where deviations cannot be readily attributed to damage due to in-situ effects [Man+11]. Influencing factors of NDT systems are reported as testing equipment and procedures, material and geometry of test specimens, and properties of the particular defect to be detected [Kur+13]. In contrast to this, SHM systems are affected by loading conditions [CFM09], temperature [Man+11], and sensor degradation [Man+11]. In [Sch+15] it is mentioned, that the assumption of independent observations is not feasible in case of continuously sampled data, because measurements performed at high acquisition rates lead to several dependent observations. Regarding the performance evaluation of a classifier, POD is understood as the true positive rate [Cai+10], which is also denoted as sensitivity, recall or detection rate (regarding object detection). In this case the probability of false alarms (false positive rate) remains unquantified. Therefore, the performance assessment of a classifier is usually based on a set of testing data with known class labels. In general, improved detection rates can only be achieved at the cost of increasing false alarm rates.

According to [RG00], only two principal ways to achieve an increase in classification performance exist. At a first sight, it seems to be possible to further increase the capabilities of an already given classification algorithm, e.g by tuning hyperparameter for specific applications. This is usually not feasible due to the individual limitations of each classification method. Therefore, a second way to improve the performance is the method of information fusion. Instead of developing individual decision systems based on one source to perform with a high reliability, the results of more than one source can be combined using information fusion methods to combine the individual advantages of different decision systems to improve the overall reliability. According to [RG00] three abstraction levels of information fusion can be divided (Figure 1.1): data fusion, feature fusion, and classifier fusion (decision fusion). By using different classifier outputs, methods on classifier level can be applied. Fusion solely relies on the subsequent combination of decisions derived by different so called base classifiers. Once the features are selected and the classifiers are tuned, classifier fusion methods only use the classifier outputs. The main advantage of this method is the exploitation of different classifier specific strengths in terms of their specific

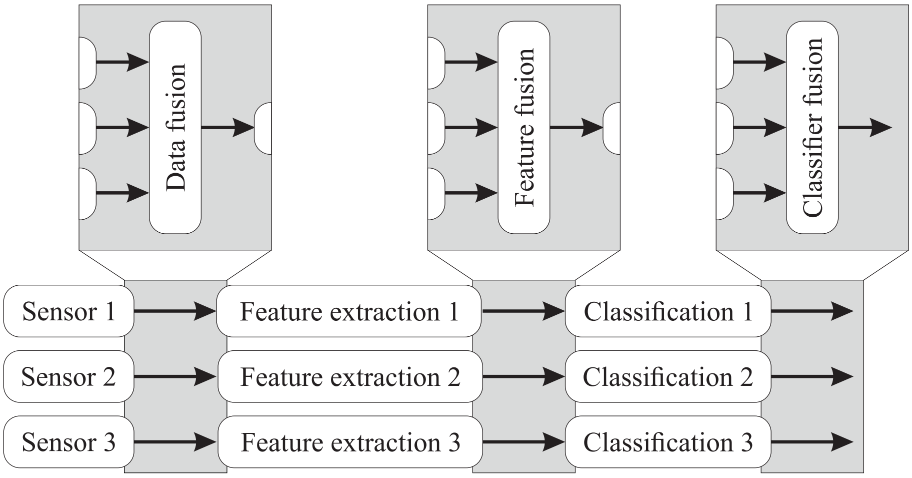

Figure 1.1: Abstraction levels of information fusion [RWS16].

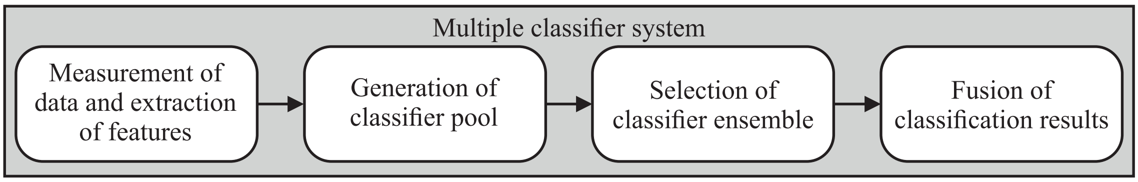
Multiple classifier system
Measurement of data and extraction of features
Generation of classifier pool
Selection of classifier ensemble
Fusion of classification results

Figure 1.2: Scheme of a multiple classifier system.

suitability for different forms of classification problems [HHS94]. The advantage of combining different classifier outputs is the independency from changes in the system behavior, so that using one individual classifier not trained with data related to the new situation can lead to a wrong decision whereas another classifier possibly not well trained, but performing better in new situations. A major drawback of decision fusion is the possibility of impeding the overall system performance by combination of several single classifiers.

Considering classifier fusion as part of a multiple classifier system, the data measurement and feature extraction, the generation of classifier pool, and the ensemble selection are done prior to the fusion itself (see Figure 1.2). The choice of suitable classifiers for combination is one of the major problems within the field of classifier fusion. To avoid the combination of classifiers with same properties and also same errors, ensemble selection strategies are used. Regardless whether suitable ensembles of classifiers have been chosen by application of static or dynamic selection methods, the parallel implementation of individual ensemble members requires further processing of results [BSO14]. For this purpose, various methods fusing individually obtained decisions have been proposed in numerous works, e.g. the Dempster-Shafer Combination [SL87], the Behavior Knowledge Space [HS93], or the Highest Rank [HHS94], which tend to improve classification performance while forming a conclusive classification result. Considering fusion methods, different characteristics can be identified. First the type of outputs generated by the individual classifiers forming an ensemble is an important property to be considered [SL00]. A further attribute distinguishing the fusion algorithms is the use of classifier outputs class-conscious or class-inherent [KBD01]. The necessity of prior training of fusion parameters divides the fusion methods into trainable and not trainable methods [Kun04].

Despite the strong researches and works in the area of improving individual classifiers, no best classifier for all data sets and applications can be identified [AS06]. This can be transferred to the application of ensemble selection [CSC18]. Considering fusion methods, the overall analysis of factors influencing the fusion performance is still an open research question. Therefore this leads to the following main research questions discussed in this thesis:

- Which selection strategy leads to the best fusion performance?

- How does the classifier performance influence the fusion performance?
- Which fusion method is suitable for which kind of data?

1.2 Thesis organization

In this thesis, the influencing factors leading to improvement of fusion performance are discussed. The thesis consists of seven chapters. Some parts of this thesis are prepared for journal papers [RS19], [RKS19], and [ARS19] or have been published in proceedings of conferences [RWS16], [RS16], [Rot^{+}17], [ARS18c], [ARS18a], and [ARS18b].

Accordingly, in the current chapter an introduction to the challenges in condition monitoring and the common solutions like classification, ensemble selection, and fusion methods are stated. Open questions are given.

The second chapter introduces the main influencing factors analyzed in this thesis. The main data characteristics, like simple measures and information theoretic measures are introduced. The performance measures used for evaluating the performance of a classifier or fusion method respectively are explained. The literature and basic ideas of static and dynamic classifier ensembles are given. Fusion method characteristics (type of classifier output level, use of classifier outputs, and the necessity of training prior to the fusion process) as well as the commonly used and in this thesis applied fusion methods are briefly introduced. The end of chapter two provides the actual developments in the field of information fusion, selection of classifiers, as well as optimizing the fusion performance.

Chapter three focuses on the analysis of the influence of ensemble selection to fusion performance. Using seven benchmark data sets and two experimental examples regarding fault diagnosis of hot rolling mills and damage detection in composites, different static and one dynamic selection strategies are applied. The selection strategies are compared and for static selection, the best classifier combination based on the performance of the individual classifiers and the optimal number of classifiers in an ensemble are identified for each data set.

In the fourth chapter the performance measure precision of one classifier (denoted as fictional classifier) in an ensemble is varied to evaluate the influence of this parameter on the overall fusion performance. The optimal precision value is determined using a non-dominated sorting algorithm. Two considerations are made: firstly the classifier is included in a specific ensemble and set as fictional classifier and secondly the addition of the fictional classifier to one other base classifier is considered. The concept of the fictional classifier is applied to four benchmark data sets with two classes, where only two parameters (precision for the two classes) are optimized. The application to fault diagnosis of hot rolling mills shows the adaptability to a four class problem, where the number of optimized values is 12. The improvement potentials as well as the properties the additional classifier should have regarding the precision value can be evaluated. Using a supervised strategy, the precision values from training are used for test to show the generalizability of the precision values from training.

The fifth chapter provides a concept using POD values or probability estimations instead of precision values as performance measure considered in the fusion process. The application to fault diagnosis of an elastic beam and to damage detection in composites show that the introduced concept provides reasonable belief values for different flaw sizes or loading conditions combining the individual performance of the detection systems.

The influence of different data and fusion method characteristics is analyzed in chapter six. The overall performance of the considered fusion methods, the performance related to the data as well as the fusion method characteristics are evaluated. In this chapter, eight different classifiers are classifying fifteen data sets from different origins. The assignments of the classifiers are fused using seven different fusion methods. Therefore the influence of the individual characteristics as well as the relation between the data characteristic and the fusion method can be emphasized.

Finally, the summary of this thesis, conclusions, and future work are outlined in chapter seven.

2 Theoretical background and literature review

In order to analyze the influences on the fusion performance, the theoretical background as well as the actual developments have to be stated.

In Chapter 2.1, all important background information is given regarding the considered data characteristics, performance measures, ensemble selection strategies, fusion method characteristics and a short explanation of the applied fusion methods. The actual developments in the field of information fusion and possibilities to optimize the performance are stated in Chapter 2.2.

The content, figures, and tables in this chapter are based on publication of [RWS16], [ARS18c], [ARS18a], and [ARS18b] and prepared for publication of [RKS19] and [ARS19].

2.1 Fundamentals of information fusion

The following chapter, contains a brief overview of data characteristics, performance measures to evaluate the reliability, different fusion method characteristics as well as common fusion methods. The chapter concludes with a detailed literature review of actual developments in the field of improving information fusion and the resulting research questions are given.

2.1.1 Data characteristics

Given the fact, that the achievable performance of the different base classifiers [MST94, WB06], and consequently also the performance of the applied fusion methods, strongly depends on the characteristics inherent to the employed sets of data, the relevant characteristics of data sets should be discussed [MST94]. In accordance to [MST94], two of the defined different main categories of data inherent characteristics, referred to as simple and information theoretic measures, are described below.

Simple measures

The group of simple measures describes the basic characteristics of a single set of data. Simple measures can be combined in terms of proportions or products etc. to generate additional information required for specific investigations [MST94]. As one of the simple measures describing the dimension of the underlying classification problem, the number of classes n_C defines the number of different groups of instances within the considered set of data. A higher number of classes results in increasing complexity of knowledge representation. The number of samples or instances n_S comprised by a single set of data increases on the one hand the computational time during training of base classifiers for a high number of samples, but also enables a more detailed generation of knowledge. Thus larger data sets should theoretically tend to improve the overall performance of classification and fusion algorithms compared to the application of smaller sets of data.

Information theoretic measures

The information theoretic measures are commonly related to the calculation of entropy denoting the mean information content. The entropy of two different relevant properties, the entropy of classes and entropy of attributes, are considered. The entropy of classes indicates the grade of evenness of the underlying class distribution. According to [MST94], the entropy is calculated using

$$H(C) = -\sum_{i} pm_i \log_2 pm_i, \tag{2.1}$$

where log_2 defines the logarithm to basis two, and pm_i represents the probability mass function for the class i (number of samples according to class i over total number of samples). For a random variable with equal probability for each of the possible values, the entropy reaches a maximum. A higher entropy of classes denotes that the number of samples according to each class is more even distributed. As the entropy represents the amount of attribute inherent information, for the entropy based on $\log_2$ the assigned unit is Bit. Considering the entropy of attributes, the formula is the same as for entropy of classes, only considering the class distribution according to one attribute. For every attribute the entropy can be calculated. If one attribute only contains samples with the same class, the entropy is 0. The more even the classes are distributed within one attribute, the entropy increases. An attribute with zero entropy however, contains no information for class discrimination, due to the non existing variation between different instances of data [MST94]. To obtain one value for the data set, the mean value of the entropies calculated for each attribute is used.

2.1.2 Performance measures

In the field of Non-destructive Testing (NDT), well-founded theory has been developed to assess the reliability of testing procedures. The Probability of Detection (POD), which is a probabilistic approach to assess the reliability of an NDT method [Kur+13], is frequently used. POD curves, which describe the likelihood that a certain flaw is detected as a function of flaw characteristic a (i.e. size or depth of a crack to be detected by an NDT-approach), can be computed directly from experimental data. In the general case, a suitable decision threshold of the sensor response $\hat{a}$ is determined. A natural lower bound of the decision threshold is given by the noise level of the measurements. As the choice of the decision threshold greatly affects the resulting POD calculations, a suitable criterion for choosing the decision threshold must be defined to account for the trade-off between minimum detectable damage size and probability of false positive detections [Kur+13]. The POD gives a general assessment of the reliability of NDT methods. The principal aim of the $a_{90/95}$ criteria is to specify a damage size, which can be detected/missed applying a specific NDT method to be evaluated, taking into account statistical variability of the sensor and measurement properties. Data used in producing POD curves are categorized by the main variables to be combined in the POD approach. These data are:

1. Hit/miss: produce binary statement or qualitative information about the existence of a flaw.
2. Flaw size vs response (a vs â): systems which also provide some quantitative measure of size of target.

A typical and useful criterion for detection at a 90 % probability of detection level with 95 % confidence level is the so-called flaw size detectability. In the derivation of the POD curve first a regression analysis of the data gathered has to be realized. The following analysis is based on the calculations given in [Dep09], [Ann17], and [GA10]. The regression equation for a line of best fit to a given data set is given by

$$y = b + mx, \tag{2.2}$$

where m is the slope and b the intercept. Here the 95 % Wald confidence bound on y is constructed by

$$y_{a=0.95} = y + 1.645\tau_y, \tag{2.3}$$

where 1.645 is z-score of 0.95 for a one-tailed standard normal distribution and τ_y the standard deviation of the regression line. The Delta method is a statistical technique used to transition from regression line to POD curve [Dep09]. The confidence bounds are computed using the covariance matrix for the mean and standard deviation POD parameters μ and σ respectively. To estimate the entries, the covariance matrix for parameters and distribution around the regression line needs to be determined. This is done using the Fisher's information matrix I. The information matrix is derived by computing the maximum likelihood function f of the standardized deviation z of the regression line values. The entries of the information matrix are calculated by the partial differential of the logarithm of the function f using the parameters of $\Theta(m, b, \tau)$ of the regression line. From

$$z_i = \frac{(y_i - (b + mx_i))}{\tau} \tag{2.4}$$

and

$$f = \prod_{i=1}^{n} \frac{1}{2\pi} e^{-\frac{1}{2}(z_i)^2} \tag{2.5}$$

the information matrix I can be computed as

$$I_{ij} = -E\left(\frac{\partial^2}{\partial\Theta_i \partial\Theta_j} \log(f)\right). \tag{2.6}$$

The inverse of the information matrix yields ϕ as

$$\phi = I^{-1} = \begin{bmatrix} \sigma_b^2 & \sigma_b\sigma_m & \sigma_b\sigma_\tau \\ \sigma_m\sigma_b & \sigma_m^2 & \sigma_m\sigma_\tau \\ \sigma_\tau\sigma_b & \sigma_\tau\sigma_m & \sigma_\tau^2 \end{bmatrix}. \tag{2.7}$$

The mean μ and standard deviation σ of the POD curve are calculated by $\mu = \frac{c-b}{m}$, where c is the decision threshold and $\sigma = \frac{\tau}{m}$. The cumulative distribution Φ is calculated as

$$\Phi(\mu, \sigma) = \frac{1}{2}\left[1 + \mathrm{erf}(\tfrac{x-\mu}{\sqrt{2}\sigma})\right]. \tag{2.8}$$

The POD function is derived as

$$POD(a) = \Phi\left[\tfrac{a-\mu}{\sigma}\right]. \tag{2.9}$$

Using this formula, the POD curve can be set up for varying flaw sizes (an example is given in Figure 2.1). The $a_{90/95}$ value denotes the maximum flaw size that can not be missed with 90 % POD at 95 % confidence level and can be taken as a performance measure.

In the past, several ideas have been reported which address different aspects to adopt POD philosophy to SHM applications. For instance, in contrast to conventional NDT the results of SHM systems are statistically not independent due to high acquisition rates [Sch$^+$15]. In this context, Schubert Kabban et al. proposed a new methodology to adopt POD procedures to provide compatibility with dependent measurement data, which is obtained from SHM systems [Sch$^+$15]. Furthermore, multiple approaches developed to assess the reliability of SHM systems are summarized by Mandache et al. [Man$^+$11]. In particular, time-based POD is proposed to address the effect of damage evolution [Man$^+$11]. It is suggested to find a formulation of POD which enables stating the probability of detecting specific defect growth within a given time interval. Multi-dimensional POD is proposed to take the effect of several in-situ effects, i.e. loading conditions, on SHM reliability into account [Man$^+$11]. This includes the computation of POD with respect to each

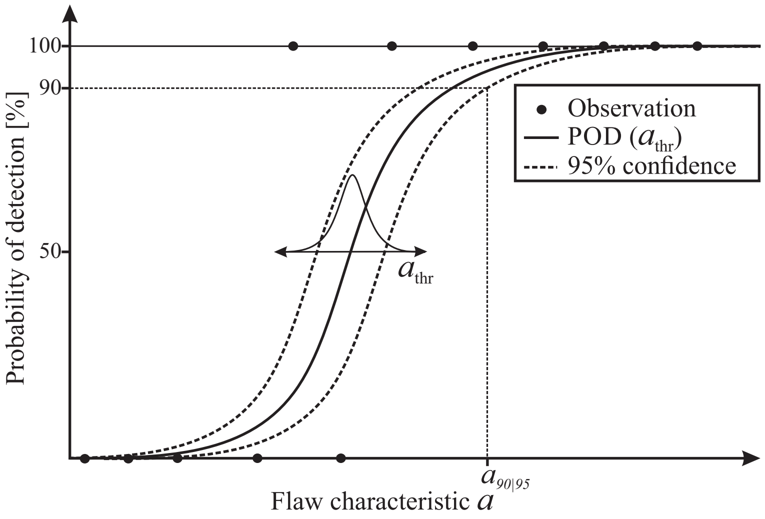

Figure 2.1: POD curve [RWS16].

influencing factor to determine the actual reliability of the SHM system in particular situations. However, the approach requires availability of quantitative information on each influencing factor. Furthermore, quantitative knowledge regarding the impact of in-situ effects on the reliability is necessary. In order to minimize the experimental effort required to determine POD, model-assisted approaches can be used [Kur$^+$13]. Cobb et al. proposed a model-assisted approach for determining POD of crack detection in aluminum specimens using in-situ ultrasonic inspection technique [CFM09]. Moreover, Eckstein et al. proposed a methodology to quantify SHM performance by using cumulative distribution functions to establish a probabilistic relationship between the detected and real damage size [EFB12]. From this method, multiple metrics of SHM performance, such as minimum detectable damage size to define a lower bound of POD as accuracy of the inspection method, and probability of false alarm are derived. However, identification of the underlying distribution functions is - particularly in context of in-situ inspection techniques, where a posteriori verification of real damage size is usually not possible - still an open issue. From the aforementioned approaches to SHM reliability assessment it is noticeable, that the common weak point is characterized by missing detailed knowledge about the impact of different factors on SHM related reliability properties.

Regarding the performance evaluation of a classifier, POD is understood as the true positive rate [Cai$^+$10], which is also denoted as sensitivity, recall or detection rate (regarding object detection). However, in this case the probability of false alarms (false positive rate) remains unquantified. Therefore, the performance assessment of a classifier is usually based on a set of testing data with known class labels. Here, the classifier output and true class labels are compared by means of a confusion matrix shown in Figure 2.2. From the confusion matrix, different scores, such as accuracy,

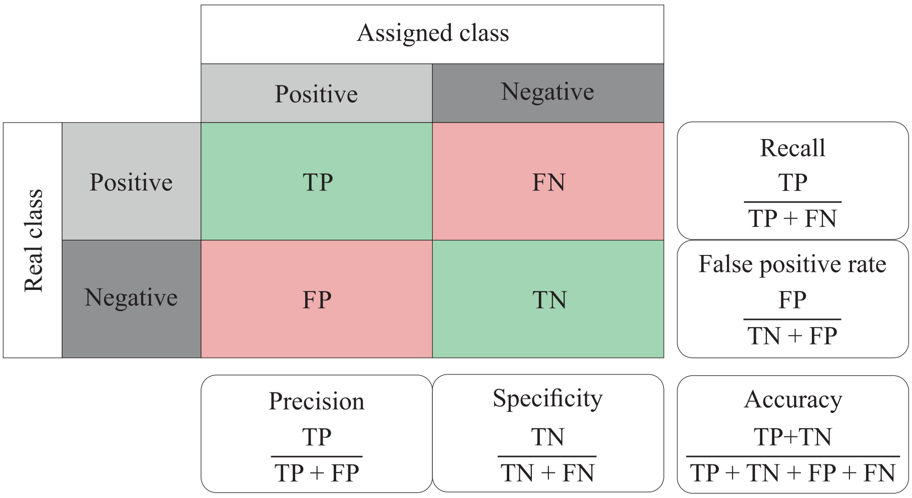
Assigned class
Positive
Negative
Real class
Positive
Negative
TP
FN
FP
TN
Recall $\frac{TP}{TP + FN}$
False positive rate $\frac{FP}{TN + FP}$
Precision $\frac{TP}{TP + FP}$
Specificity $\frac{TN}{TN + FN}$
Accuracy $\frac{TP+TN}{TP + TN + FP + FN}$

Figure 2.2: Confusion matrix and related performance measures [RWS16].

precision, specificity, recall, and false positive rate are extracted. The difference between accuracy and recall is that accuracy considers all correct classifications in relation to all classification assignments, recall just takes the positive classifications (e.g. object is present) into account. Calculating the correct positive classifications over all positive assignments, the precision value is a measure of the reliability of the assignments of one classifier. In general, improved detection rates can only be achieved at the cost of increasing false alarm rates. The principle relationship between detection and false alarm rate is described using the ROC curve [MB13], which compares the detection and false alarm rate of a classifier. In contrast to POD the ROC curve provides a suitable method to assess the overall performance of a classifier [Cai+10, WY07].

To compute a measure for improvement obtained by fusion of all classifiers, for each set of data as well as fusion method, the accuracy of the best individual classifier $\text{Accuracy}_{\text{best}}$ is compared to the accuracy of the fused ensemble itself $\text{Accuracy}_{\text{fused}}$. The partitioning of every data set is exactly the same for all of the considered classifiers and fusion methods, the comparison of accuracy is conducted in accordance to the matched sample approach suggested by [Won15]. Therefore, the gain of accuracy ΔAcc can be computed by

$$\Delta Acc = \text{Accuracy}_{\text{fused}} - \text{Accuracy}_{\text{best}}. \tag{2.10}$$

2.1.3 Ensemble selection strategies

To avoid the combination of classifiers with low accuracy and high dependency in comparison to other classifiers, ensemble selection methods are used. The ensemble selection methods are divided into two strategies, the Static Classifier Ensemble (SCE) and the Dynamic Classifier Ensemble (DCE) strategy.

Static Classifier Ensemble (SCE)

The SCE method obtains selection regarding all unseen patterns based on validation errors during training phase. This approach selects an ensemble of suitable classifiers for combination [ME15, BSO14]. A representative example of such an SCE method is the so called test and select approach as proposed in [Sha$^+$00]. Basically this method is comparing different subsets of available classifiers regarding accuracy on a validation set, accordingly selecting the best performing subset i.e. ensemble and further testing the performance on held aside test data. Ensemble selection methods resort to the usage of fusion methods [Yan11, Sha$^+$00].

Dynamic Classifier Ensemble (DCE)

Based on the concept of selecting classifiers taking into account their performance on validation data during training phase and considering different parameters of the current sample during classification, DCE methods aim to select groups of appropriate classifiers instead of selecting only a single best entity. Thus for example avoiding potential deficiencies of single classifiers regarding individual feature spaces is accomplished [ME15], hence raising classification performance while facing the varying requirements of individual samples. As pointed out by [KSB08] specific DCE approaches are in certain cases superior to SCE approaches. Due to the selection of multiple classifiers, DCE methods resort to the subsequent usage of fusion methods [BSO14, KSB08].

An exemplary implementation representing DCE methods is the so called DCE-clustering method proposed by [Soa$^+$06]. The reasoning behind this method is to primarily cluster the validation data by application of k-means algorithm [Kun04] and subsequently assigning proper ensembles for each group of instances. Assignment is performed based on ranking the base classifiers according their accuracy and diversity by evaluating pairwise diversity measures such as the double fault measure [GR01]. During implementation each current sample is assigned to its nearest centered cluster (euclidean distance), where a predefined number of the most accurate and diverse classifiers is chosen [Soa$^+$06].

A modification of the former procedure is the so called DCE-k-Nearest-Neighbor method. Instead of clustering validation data, this approach pursues calculation of sample similarity based on configuration of a k-Nearest-Neighbor classifier obtained from validation data. During application, each current sample is assigned to its k nearest neighbors, according to which the most accurate and diverse classifiers are chosen similar to the prior mentioned strategy [Soa$^+$06]. A more elaborated discussion of these and several other dynamic selection approaches such as k-nearest oracle or randomized reference classifier is e.g. given in [CSC18], [ME15] and [KSB08].

2.1.4 Fusion method characteristics

Prior to elaborating different algorithms for decision combination, the attributes and requirements with respect to different methods of classifier fusion should be discussed. This chapter therefore is further subdivided into the description of different output levels, the different use of these outputs, and the necessity of training of the inherent parameters of fusion methods [Kun04].

Type of classifier output levels

Considering the utilized type of classifier output levels as an attribute of the considered fusion method, different definitions exist. According to [Bez$^+$06] the output levels can be divided into possibilistic, probabilistic, and crisp labels. Using this notation, the possibilistic labels can be interpreted as possibility of class membership [And$^+$10], the probabilistic labels as posterior probabilities [DT98], and the crisp labels denote an assignment of classifier for a single class. However, as the final output obtained by a specific classifier is not limited to a single class label a more practical categorization is proposed in [XKS92], which divides the different levels of classifier outputs into abstract, rank, and measurement level. The abstract level is equivalent to the crisp labels, but both possibilistic and probabilistic labels belong to measurement level. The rank level represents an ordered subset including the most plausible classes. Considering the amount of information inherent to each of the depicted categories, the least amount of information is provided on the abstract level, given the fact that only a single class label is generated without declaration of certainty or the information of alternative class labels [Kun04]. In contrast, information based on the measurement level provide the highest amount of useful data, due to direct propagation of possibilistic or probabilistic labels as final classifier output [RG00]. However, caused by different mathematical backgrounds of classifiers, the application of information on measurement level often requires further normalization of results to ensure reasonable combination [HHS94]. Transformation of an output with high level information, such as the measurement level, into an output of lower information density is always possible [RG00]. This can be justified by the fact, that for the generation of an ordered set of possible class labels (i.e rank level) as well as for the assignment of a single class label (i.e. abstract level) the amount of higher level information is merely reduced [XKS92]. The transformation of output levels in direction of a higher information density is only possible, if additional data resulting from training of the individual classifier are available [RG00]. The similarity of all definitions for output labels is the clear distinction between abstract level (related to crisp labels) and soft level (related to rank and measurement level as well as possibilistic and probabilistic labels).

Use of classifier outputs

The base classifiers assign either a support value for each class (rank, possibilistic, or probabilistic) or a single class (support value equal to 1 for the supported class, all other values are zero) as output. Depending on how the fusion method is using the information of classifier outputs for final decision, the fusion methods are either denoted as class-conscious or class-indifferent [KBD01]. The class-conscious fusion methods only use the support values of the considered class, neglecting the support values of the other classes. The class-indifferent methods include all support values into the decision process. According to [KBD01] class-conscious methods consider the class context, but neglect some information given by the classifiers, whereas the class-indifferent methods use all information, but ignore the context.

Training of fusion method parameters

According to [Dui02] fusion methods can be divided into trainable and fixed methods. The difference is the necessity of an additional training prior to the fusion process to set method-specific parameters. These different parameters could be weights associated with specific base classifiers, as done during application of Logistic Regression [HHS94]. Another example is the training of conditional class probabilities computed in case of applying Bayes Belief Integration [XKS92]. Approaches without any training of parameters are for example the Majority Voting or the Borda Count method to be used directly after classification. Considering fusion methods using training of parameters, an additional amount of data samples is required. As stated in [Sue90] and [Dui02], the number of training data samples as well as the use of the same data for the training of base classifiers and fusion methods is significant in relation to the fusion results.

2.1.5 Common information fusion methods

To explain the different fusion methods, the fundamental problem and the related variables have to be defined.

Each sample x of a data set Z with the total number of samples n_S is related to one class c_i, where $i = 1, \ldots, n_C$ with n_C as the total number of classes of Z. The classification is done by n_K classifiers, where each classifier e_k assigns an output for each sample x. Whether the output of the classifier is on abstract level, a ranking or a measurement output, the generated classifier output is denoted as $q_k(x)$, $r_k(x)$ or $s_k(x)$ respectively. While q_k is a scalar, the ranking output contains n_C elements with $r_k(x) = [r_{k,1}(x) \ldots r_{k,n_C}(x)]^{\mathrm{T}}$, where $r_{k,i}(x)$ is the assigned rank for class c_i. The measurement outputs also consists of n_C elements with $s_k(x) = [s_{k,1}(x) \ldots s_{k,n_C}(x)]^{\mathrm{T}}$, where $s_{k,i}(x)$ is a support value for specific class c_i. Using fusion, a new output $\hat{q}(x)$ for abstract level, $\hat{r}(x) = [\hat{r}_1(x) \ldots \hat{r}_{n_C}(x)]^{\mathrm{T}}$ for ranking labels, and $\hat{s}(x) = [\hat{s}_1(x) \ldots \hat{s}_{n_C}(x)]^{\mathrm{T}}$ for measurements labels is produced.

In the following, nine commonly used fusion methods are shortly explained.

Majority Voting

The Majority Voting (MV) is based on the simple majority rule, where a selection or decision is made based on the number of votes for each alternative solution. Similar to this idea, during fusion using MV, the final decision is based on the number of classifiers assigning a specific class to a sample. The assignment of the classifier has to be on abstract level. Therefore to provide for whether classifier e_k assigns a class $q_k(x)$ to a given sample x, the binary characteristic function for considered class c_i is introduced, where

$$b_k(x \in c_i) = \begin{cases} 1, & \text{if } q_k(x) = c_i \\ 0, & \text{otherwise.} \end{cases} \tag{2.11}$$

For each class, the binary characteristic functions for all n_K classifiers are added to the number of votes for each class c_i using

$$b(x \in c_i) = \sum_{k=1}^{n_K} b_k(x \in c_i). \tag{2.12}$$

Consequently, with respect to the current pattern x, the number of votes for each class determines the choice of a final class label by application of

$$\hat{q}(x) = \begin{cases} c_i, & \text{if } b(x \in c_i) = \max\left\{b(x \in c_1), \ldots, b(x \in c_{n_C})\right\} \geq \alpha \cdot n_K + f_t(x) \\ 0, & \text{otherwise,} \end{cases} \tag{2.13}$$

where $\alpha \in (0, 1]$ represents a parameter that affects the required number of votes for a specific class label to be considered as the final result $\hat{q}(x)$. Moreover the additive function $f_t(x)$ usually considers the exception of receiving the same number of votes for the top classes or the frequent case of a slight difference of voting [XKS92]. Equation 2.13 describes a broad variety of voting methods, each defined by its distinct implementations of parameters [RG00, XKS92], the specific version termed as MV is obtained by defining $\alpha = 0.5$ and $f_t(x) = 0$ [Sue90, RG00].

Weighted Voting

The fusion method Weighted Voting (WV) is similar to MV, except of the weighting, which is considered during the fusion process. Each vote is weighted with a given value (previously set or trained). If the classifier's output is a confidence value, this can be used as weighting value for each assignment. Else the output is a crisp value on abstract level, the weights can be defined as precision value p_i^k for each class c_i and each classifier e_k using a training data set. The weighted characteristic function for each classifier and class can be calculated using

$$w_k(x \in c_i) = \begin{cases} p_i^k, & \text{if } q_k(x) = c_i \\ 0, & \text{otherwise.} \end{cases} \tag{2.14}$$

Similar to MV, the weighting characteristic functions are added for each class by

$$w(x \in c_i) = \sum_{k=1}^{n_K} w_k(x \in c_i). \tag{2.15}$$

By comparing the added precisions for each class, the class with the maximum value is set as final decision

$$\hat{q}(x) = \{c_i \mid w(x \in c_i) = \max\big\{w(x \in c_1), \ldots, w(x \in c_{n_C})\big\}\} \tag{2.16}$$

Different to MV, the maximum number of added votes do not have to reach a specific parameter α, only the class with the maximum sum w is denoted as $\hat{q}(x)$. Ties are solved randomly.

This fusion method contains the advantage of MV regarding the simple implementation, but combining this with the knowledge about performance of individual classifiers from training [KR14].

Bayesian Combination Rule (BCR)

The Bayesian Combination Rule (BCR) also known as Bayes Belief Integration or Bayesian Belief Method is a well known and commonly used fusion technique based on conditional probability.

To set up the conditional probabilities of each classifier for each class, first the confusion matrix has to be calculated. The confusion matrix C^k for each classifier e_k is defined as

$$C^k = \begin{bmatrix} C_{11}^k & C_{12}^k & \cdots & C_{1n_C}^k \\ C_{21}^k & C_{22}^k & \cdots & C_{2n_C}^k \\ \vdots & \vdots & \ddots & \vdots \\ C_{n_C 1}^k & C_{n_C 2}^k & \cdots & C_{n_C n_C}^k, \end{bmatrix}, \tag{2.17}$$

where $i, j = 1, ..., n_C$ with n_C as the number of classes. The element C^k_{ij} is the number of samples, where the classifier e_k has assigned class c_j and the actual class of the sample is c_i.

Using the elements of the confusion matrix the probability, that sample x belongs to class c_i, if the classifier e_k assigns x to class c_j can be calculated using

$$P^k_{ij} = P(x \in c_i \mid q_k(x) = c_j) = \frac{C^k_{ij}}{\sum_{i=1}^{n_C} C^k_{ij}}. \tag{2.18}$$

For each classifier e_k the probability matrix P^k is set with

$$P^k = \begin{bmatrix} P^k_{11} & P^k_{12} & \dots & P^k_{1n_C} \\ P^k_{21} & P^k_{22} & \dots & P^k_{2n_C} \\ \vdots & \vdots & \ddots & \vdots \\ P^k_{n_C1} & P^k_{n_C2} & \dots & P^k_{n_Cn_C} \end{bmatrix}. \tag{2.19}$$

The diagonal values $(i = j)$ are the same as the precision value p_i for this class. Based on the probability matrix of each classifier, a combined belief value $bel(i)$ for each class i is determined for each sample with the formula

$$bel(x \in c_i) = \frac{\prod_{k=1}^{n_K} P^{n_K}_{iq_k}}{\sum_{i=1}^{n_C} \prod_{k=1}^{n_K} P^{n_K}_{iq_k}}, \tag{2.20}$$

where q_k is the assigned class of classifier e_k for the considered sample x. The maximum of the belief values is used to make a decision $\hat{q}(x)$ for one of the classes.

Behavior Knowledge Space

The Behavior Knowledge Space (BKS) method [HS93], uses the specific combination of classifier labels from a training data set to denote a most probable class for an unknown sample generating a new combination of classifier labels. While considering the individual assignments $q_k(x)$ of the n_K different classifiers and the n_C possible class assignments, the $n_C^{n_K}$ possible combinations describes the Behavior Knowledge Space. Every specific combination of labels is called a unit of the BKS, and further denoted as $BKS(q_1(x), \dots, q_{n_K}(x)) = BKS_l$, where $l = 1, \dots, n_C^{n_K}$. During training, for each unit BKS_l the total number of samples n_l as well as the number of samples belonging to one class $n_l(c_i)$ is counted. The best representative class for each unit BKS_l is denoted by

$$c_l = \{c_i \mid n_l(c_i) = \max\{n_l(c_1), \dots, n_l(c_2)\}\}. \tag{2.21}$$

Once the best representative classes are determined, the combination of assignments for the unknown sample x is related to the unit and the final decision is made by

$$\hat{q}(x) = \begin{cases} c_l, & \text{if } n_l > 0 \text{ and } \dfrac{n_l(c_l)}{n_l} \geq \lambda \\ 0, & \text{otherwise.} \end{cases} \tag{2.22}$$

where $\lambda \in [0, 1]$ is a parameter controlling the degree of confidence in the generated decision [RG00], while the original chosen value (as proposed in [HS93]) is $\lambda = 0.5$.

Given the case that the number of training samples assigned to a single unit equals zero, and therefore there exists no knowledge on possible class labels, the final decision is reached at random from the set $\{c_1, \ldots, c_{n_C}\}$ [KBD01]. Furthermore if ties between different classes exist within a unit, these ties are broken randomly [KBD01].

Dempster-Shafer Combination

Developed by A. P. Dempster [Dem67] and extended by G. Shafer [Sha76], the Dempster-Shafer Theory (DST) evaluates how to deal with uncertainties. The Dempster-Shafer Combination (DSC) is used to combine different sources with specific assignments to one assignment considering also uncertainties.

The frame of discernment (here: set of all classes c_i) $\Theta = \{c_1, c_2, \ldots, c_{n_C}\}$ is extended to the power set 2^Θ, which contains all possible subsets of Θ including the empty set $\emptyset$.

For each element of the power set 2^Θ a basic assignment with $m : 2^\Theta \rightarrow [0, 1]$ is assumed, where $m(\emptyset) = 0$ and $\sum_{A \subseteq \Theta} m(A) = 1$. To combine the basic assignments of several classifiers, the joint mass

$$m(A) = \frac{\sum_{B \cap C \cap \ldots \cap X = A \neq \emptyset} m_1(B) m_2(C) \ldots m_K(X)}{1 - \sum_{B \cap C \cap \ldots \cap X = \emptyset} m_1(B) m_2(C) \ldots m_K(X)} \tag{2.23}$$

is calculated for each class, where A, B, C, $\ldots$, X are the elements of the power set 2^Θ. The maximum joint mass assigns the final decision $\hat{q}(x)$ to one of the elements from the power set.

Highest Rank

Using the Highest Rank (HR) fusion method, the required outputs of classifiers have to be rankings $r_k(x)$ with highest (= best) rank is equal to one. For each class c_i

with $i = 1, \ldots, n_C$, the minimum rank from all classifiers $r_{k,i}$ with $k = 1, \ldots, n_K$ is determined according to

$$\hat{r}_i(x) = \min \left\{ r_{1,i}(x), \ldots, r_{n_K,i} \right\} \tag{2.24}$$

Subsequently, the classes are ordered according to the new rank $\hat{r}_i(x)$. The final decision for one class can be obtained by just using the class with the highest rank. Given the case that different class labels would receive an identical position in the final ranking, the conflicts are broken by random.

Borda Count

The fusion method Borda Count (BC) is an extension of MV, using rankings instead of specific class labels. Therefor, the so-called borda count bc_i is calculated for each class c_i using the total number of classes ranked below the considered class and add this number for all classifiers. Thus the borda count is calculated using

$$bc_i = \sum_{k=1}^{n_K} (n_C - r_{k,i}(x)), \tag{2.25}$$

where $r_{k,i}(x)$ is the rank assigned by the current analyzed classifier k under consideration of pattern x. Using the resulting values in descending order, the final ranking can be set up according to the rank of the border count of the individual classes. Potential ties arising during the development of the final ranking $\hat{r}(x)$ are for example arbitrarily broken [Sch+04]. As an exception the application of this method in case of a two class problem, is equal to the application of the aforementioned MV procedure [RG00].

Logistic Regression

Using the fusion method Logistic Regression (LR), the former notation of ranks regarding the definition of highest rank equals one is invalid. Now the highest rank has to be denoted with the highest possible value n_C. Using this notation, a new score vector $v_i(x) = [v_{1,i}(x) \ldots v_{n_K,i}(x)]^{\mathrm{T}}$ for each class c_i is defined, with $v_{k,i}(x)$ as assigned rank of class c_i from classifier e_k. During training, the true class is known and denoted by the response variable $Y = 1$. For all samples of data sets, the score vectors for each sample and class are determined as well as the related Y value. The empirical probability, that a specific score vector v with a specific rank notation is denoting the true class can be calculated by

$$P(Y = 1 \mid v) = \frac{N_{(Y=1)}}{N}, \tag{2.26}$$

where N denotes the number of classes receiving v and $N_{(Y=1)}$ the number of true classes at v. The so-called empirical logits can be calculated by

$$L(v) = \log\left(\frac{P(Y = 1 \mid v)}{1 - P(Y = 1 \mid v)}\right). \tag{2.27}$$

Using the logistic response function

$$P(Y = 1 \mid v) = \frac{\exp(\alpha + \beta_1 v_1 + \ \ldots \ + \beta_{n_K} v_{n_k})}{1 + \exp(\alpha + \beta_1 v_1 + \ \ldots \ + \beta_{n_K} v_{n_k})}, \tag{2.28}$$

and the formula for logits

$$L(v) = \log\left(\frac{P(Y = 1 \mid v)}{1 - P(Y = 1 \mid v)}\right) = \alpha + \beta_1 v_1 + \ \ldots \ + \beta_{n_K} v_{n_k}, \tag{2.29}$$

the constant model parameters α and $\beta = [\beta_1 \ldots \beta_{n_K}]$ can be estimated using methods related to linear regression. According to [HHS94], methods based on maximum likelihood or weighted least-squares can be used.

For each test sample the probability $P(Y = 1 \mid v_i(x))$ can be calculated and the class with the highest probability, that this class is the true class ($Y = 1$) is set as final class label. As there is often only a finite amount of data for training, a single Θ can be obtained for all classes, thus reducing the required amount of training data [HHS94].

Fuzzy Integral

The application of fuzzy integrals (FI) within the field of classifier fusion, is often interpreted as searching for the maximum agreement between the individual classifiers decisions and a specific generated fuzzy measure for each class [CK95]. To realize the FI method, the classifier outputs have to be on measurement level $s_k(x)$. The fuzzy measure for each class has to be a value between 0 and 1. Finding these measures is the key problem which is solved with training data. First, the fuzzy densities g^i for each class are determined, e.g. by using the precision value p_i^k [CK95] or using $p_i^k/2$ [Wan$^+$98]. Given these different fuzzy densities, derived during training phase, the computation of λ is obtained by solving equation

$$\lambda + 1 = \prod_{i=1}^{n_C} \left(1 + \lambda g^i\right). \tag{2.30}$$

Using the Sugeno Fuzzy Integral method (solely considered within this work), for each class c_i the support values of all classifiers $s_{1,i}(x), \ldots \, s_{n_K,i}(x)$ are sorted in descending order with the index j. The corresponding fuzzy densities are calculated by

$$g_{j,i} = \begin{cases} g^i, & \text{if} \quad j = 1 \\ g^i + g_{j-1,i} + \lambda g^i g_{j-1,i}, & \text{if} \quad 1 < j \leq n_K. \end{cases} \tag{2.31}$$

For each class, the final degree of support $\omega_i(x)$ for class c_i is denoted as the support value $s_{k,i}(x)$, for which the fuzzy density $g_{j,i}$ is the lowest value in this class c_i. The fuzzy integral e is calculated using

$$e = \max\left\{\omega_1(x), \ldots, \omega_{n_C}\right\}. \tag{2.32}$$

The final decision for one class is the corresponding class for e.

2.2 Actual developments

Using fusion methods the goal to combine the individual advantages of classifiers to improve the overall performance can be realized [Tul+08].

From literature several contributions [Bil15, FL09, WTG09, Man+15], focusing the improvement of the performance of the decision support system by using simple fusion methods like voting methods (e.g. Weighted Voting), sum rules, or averaging probabilities are known. Although using similar fusion methods, the application field varies, e.g. the decisions are in diverse context. Also like diagnostics of an accelerometer [Bil15], face recognition [FL09], classifying students' learning in online learning process [WTG09], or land cover classification [Man+15].

In the application field of fault diagnosis, the improvement of classification results using fusion is still challenging. In [QZG16] classification results are combined using the Bayesian Combination Rule. In the cited contribution the accuracy of fused classification results for fault diagnosis of gearbox and locomotive bearings is higher than the individual classification results. Another example using BCR for bearing fault diagnosis is given in [XKH11], here the performance of a single classifier is outperformed using the proposed fusion algorithm. The Dempster-Shafer Combination based on the Dempster-Shafer Theory is another fusion method on the measurement level. For fault diagnosis, the Dempster-Shafer Combination is used to combine neural network classifier outputs as given in [HB11] for fault diagnosis of induction motor, of railway track circuits [Ouk+10], or for spark plug fault recognition applied on different data sources [Moo+15]. In all these contributions, the decision support system is improved in performance using the Dempster-Shafer Combination.

A detailed comparison of the results achievable with DSC for 13 different NASA data sets is done in [PG11]. The NASA data sets classified by WEKA (an open source collection of machine learning algorithms for data mining tasks [Hal$^+$09]) are combined to prove the possibility of improvement using the DSC [PG11]. Resulting from the study in [PG11], the DSC is able to achieve higher probability of detection than individual classifiers, but not in all cases or for all data sets. In [ZD11] several classifier fusion methods are compared to analyze the performance to Structural Health Monitoring systems. First, nine fusion methods with different combinations of classifiers are compared using synthetic data. The DSC and the fuzzy logic type 2 algorithms were claimed to have best results concerning the correct classification rate. To validate the results, DSC and fuzzy logic type 2 are compared to fuzzy logic type 1 and majority voting using experimental data of structural damages of an aluminum plate.

In [Wan$^+$14] the fusion method Fuzzy Integral is improved by using upper integrals regarding the accuracy of classifier combinations. The results are produced using 14 data sets from the UCR Time Series Classification Repository [Che$^+$14]. The proposed results indicate that the accuracy using upper integrals are not lower than the individual accuracies of the classifiers, so not outperforming the individual performance. In [HVY11] the data sets provided by the UCR Time Series Classification Repository are also used to test the proposed extension of the DSC with the decision template. Solutions for the problem of dependencies between the fused classifiers are given in [QMD11] and [MYL13].

The effect of the training sample size on the performance of fusion methods is determined in [MZB12]. In this study [MZB12], 10 different classifiers are combined using 13 different fusion methods (e.g. mean value, maximum value, majority voting, BCR, decision tree) and also a two-level fusion is considered. Only one real data set from a hand gesture recognition problem is used. From [ZD11] it is known that two-level fusion using different kinds of fusion methods produces better results as one-level fusion and performs a slightly better using larger training sample sizes. In [Tam$^+$11] a three level fusion is realized using 5 classifiers and a genetic algorithm to find the optimal level structure. For two data sets with the background from physiological monitoring in Body Area Networks the applied approach based on genetic algorithm presented in [Tam$^+$11] increases the accuracy. To optimize the fusion performance, optimization techniques can be used to find optimal parameter within the fusion method. In [NL09] the Correspondence Analysis (CA) is used to combine the classifier assignments. Using validation data, the parameters of the CA are optimized by a genetic algorithm. This results in better performances of six benchmark data sets compared to other methods without optimization. Also in [Ngu$^+$18] an optimization technique is used to find optimal parameter used during the fusion progress by analyzing validation data. The dependence of error rate on this parameter is shown in [Ngu$^+$18], where the optimal parameter varies for the considered 21 different benchmark data sets.

From the given literature it can be concluded, that fusion methods are widely used in several application fields. The applied fusion methods can improve the accuracy of a decision support system. Optimization techniques can help to find optimal parameters related to the applied fusion method or the input to the fusion method (e.g. selected classifiers or features in an ensemble). The applied fusion methods can improve the accuracy, but as shown in several literature, an improvement using fusion methods is not always possible for all data sets.

Instead of using optimization of fusion methods, ensemble selection methods are designed to avoid the combination of classifiers with low accuracy or high dependency in comparison to other classifiers. The achievable performance of fusion methods mainly relies on the selection of the most diverse and accurate single base classifiers [HHS94]. According to [AS06] no best classifier for all problems exist and the individual performance of classification methods itself also depends on several characteristics inherent to the classified set of data [WB06]. A combination of SCE using genetic algorithm and DCE using k-nearest neighbor is used to establish a hybrid methodology in [ME15]. Considering DCE strategies, the competence measurement is defined differently, e.g. as classification confidence [Li+13], overall local accuracy [Vri+15, NN12], or the local class accuracy [Vri+15]. Also a so-called diversity measurement can be used for selection of the most independent classifiers for ensemble. A detailed analysis of the influence of three different diversity measurements using 14 data sets was done in [FCX15]. The ensemble selection is used to improve the accuracy by modifying the input to the fusion method, but the dependency of the results on the fusion methods is not considered. Instead of using k-nearest neighbor, clustering techniques can be used to find the region of interest, like in [Lin+14]. As described in [CSC18] the DCE strategies vary in the definition of the region of competence (like clustering, K-Nearest neighbor, Potential function model, or decision space), the selection criteria (e.g. accuracy, oracle, diversity), and the selection mechanism (single classifier or ensemble selection). Although a comparative study is carried out using 18 selection techniques with 30 benchmark data sets, no ideal selection technique can be found. Optimization algorithms are also used to get the optimal static classifier ensemble. In [Ngu+14] the fusion algorithm as well as the classifier ensemble including the features are optimized to maximize the number of correct classified samples and to minimize the number of selected features and classifiers. The results show, that a reduced number of selected features and classifiers in combination with a simple fusion method (sum rule) can result in comparable accuracy values than a complex fusion algorithm. The optimization of weights taken as input of the DSC fusion method is proposed in [Liu+18]. The optimization is based on evidential reasoning and minimizes the distance between the combination result of classifiers and the true class label.

In general, no conclusive statement of the best fusion or selection method can be given, independent from the statistical base of the individual fusion approaches. From the literature it can be concluded that the results are strongly affected by application data, fusion method and selection strategy. In the literature individual fusion methods are optimized for specific applications or only a few parameters are considered. Therefore the concrete consideration of influencing factors and their specific effect on the overall fusion performance is still an open question. In detail, all parts (data characteristics, classifier properties, selection strategies, and fusion method characteristics as well as the methods themselves) of the introduced multiple classifier system should be considered. The related open research questions can be detailed as follows. Regarding the data characteristics, the analysis of the properties leading to an improvement in fusion as well as a recommendation of a fusion method for a specific kind of data set should be conducted. By varying the classifier properties considered in the fusion process, a statement about the properties an additional classifier should have or which classifier parameters could be changed to improve the overall fusion performance can be stated. The static and dynamic selection strategies could be compared and analyzed with regard to the optimal number of classifiers in an ensemble or which properties the classifiers should have to optimize the fusion result. Different fusion methods with different characteristics should be compared to show effects on the fusion performance. An overview of the research questions related to the parts of the MCS is shown in Figure 2.3.

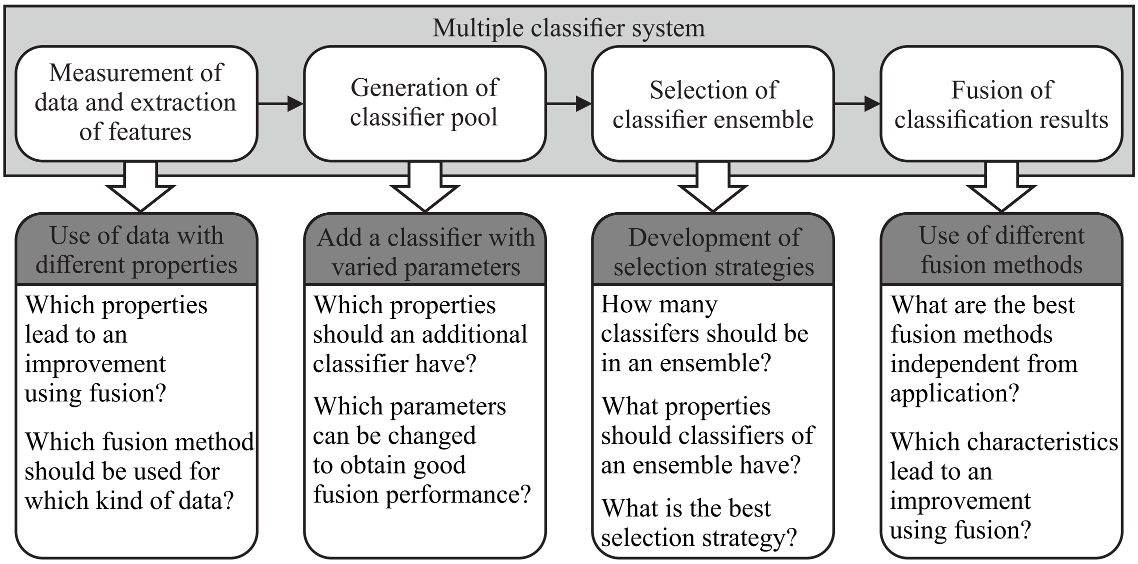
Multiple classifier system
Measurement of data and extraction of features
Generation of classifier pool
Selection of classifier ensemble
Fusion of classification results
Use of data with different properties
Which properties lead to an improvement using fusion?
Which fusion method should be used for which kind of data?
Add a classifier with varied parameters
Which properties should an additional classifier have?
Which parameters can be changed to obtain good fusion performance?
Development of selection strategies
How many classifers should be in an ensemble?
What properties should classifiers of an ensemble have?
What is the best selection strategy?
Use of different fusion methods
What are the best fusion methods independent from application?
Which characteristics lead to an improvement using fusion?

Figure 2.3: Scheme of a multiple classifier system and related research questions answered in this work.

3 Comparison of fusion performance using ensemble selection

To examine the effects from data structure to resulting accuracy using different ensemble selection strategies and fusion methods in this chapter some basic research and combination calculations are realized. A detailed comparison of three different fusion methods (WV, BCR, and DSC) using different selection strategies (SCE and DCE) is given. As performance measure, the overall accuracy of the individual classifier and the precision of each class for each classifier is used. Using a set of benchmark data as well as two experimental data sets the results are numerically analyzed concerning the number and quality of classifiers to be combined. Based on this (general) example, questions about a suitable combination of ensemble selection and fusion method can be answered.

In Chapter 3.1, the concept using SCE and DCE for the applied data is explained. The numerical results are analyzed and compared in Chapter 3.2. In Chapter 3.3 experimental results using the introduced selection strategies are discussed for two different examples for condition monitoring. Conclusions are given in Chapter 3.4.

The content, figures, and tables in this chapter are based on publication of [RS16] and [Rot^{+}17]. Part of the content, figures, and tables in this chapter are modified after previous publication.

3.1 Concept of applied selection strategies

For selecting classifiers, two strategies are considered. The first one is the static selection of classifiers (SCE) based on the overall accuracy of the individual classifiers. The whole data sets are used to evaluate the different ensembles. The second strategy is the dynamic selection of classifiers (DCE). The data sets are divided into validation and test data sets (two thirds to one third and one third to two thirds).

3.1.1 Static Classifier Ensemble strategy

Using an SCE strategy, the validation data is classified by all available classifiers and the accuracy is calculated individual for each classifier (see Figure 3.1). Once the classifier ensemble is set, the test data can be classified by the classifiers in the specific ensemble. The results are fused to get one final decision.

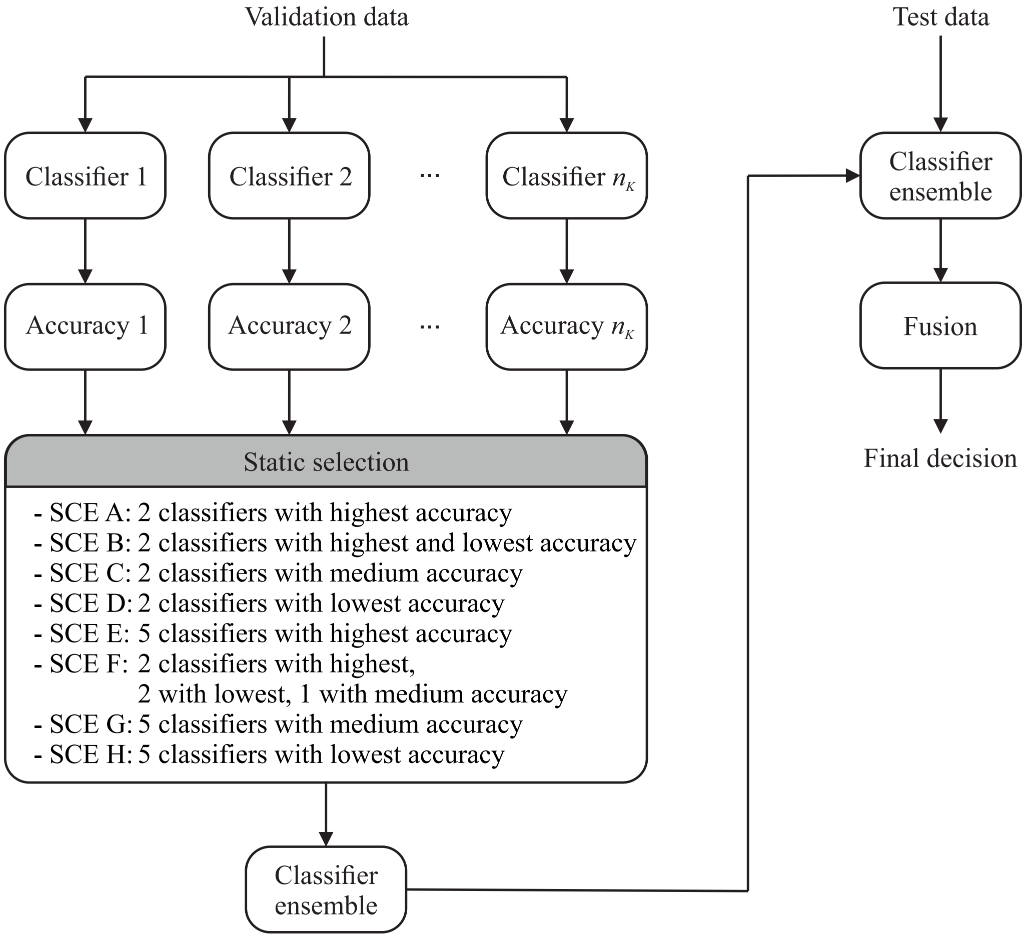
Validation data
Test data
Classifier 1
Classifier 2
···
Classifier n_K
Classifier ensemble
Accuracy 1
Accuracy 2
···
Accuracy n_K
Fusion
Final decision
Static selection
- SCE A: 2 classifiers with highest accuracy
- SCE B: 2 classifiers with highest and lowest accuracy
- SCE C: 2 classifiers with medium accuracy
- SCE D: 2 classifiers with lowest accuracy
- SCE E: 5 classifiers with highest accuracy
- SCE F: 2 classifiers with highest, 2 with lowest, 1 with medium accuracy
- SCE G: 5 classifiers with medium accuracy
- SCE H: 5 classifiers with lowest accuracy
Classifier ensemble

Figure 3.1: Concept of applied Static Classifier Ensemble strategy.

3.1.2 Dynamic Classifier Ensemble strategy

Independent from the accuracy, the individual classifiers can be better or worse for individual classes. The precision of each class p_i for each classifier is a measure of the reliability of the assumptions. To ensure, that the classifiers with the highest precision are chosen to build an ensemble, the classifiers are selected for each data set using their precision for each class to get the DCE.

The concept of the applied DCE strategy is depicted in Figure 3.2. During the validation, the precision for each class and classifier is calculated using validation data with known samples. For each class the four classifiers with the highest precision are set as suitable classifiers, so that each classifier is assigned to a set of classes, for which the precision value was in the top four of the class. The number of classes can vary and a classifier can be not suitable for any class also.

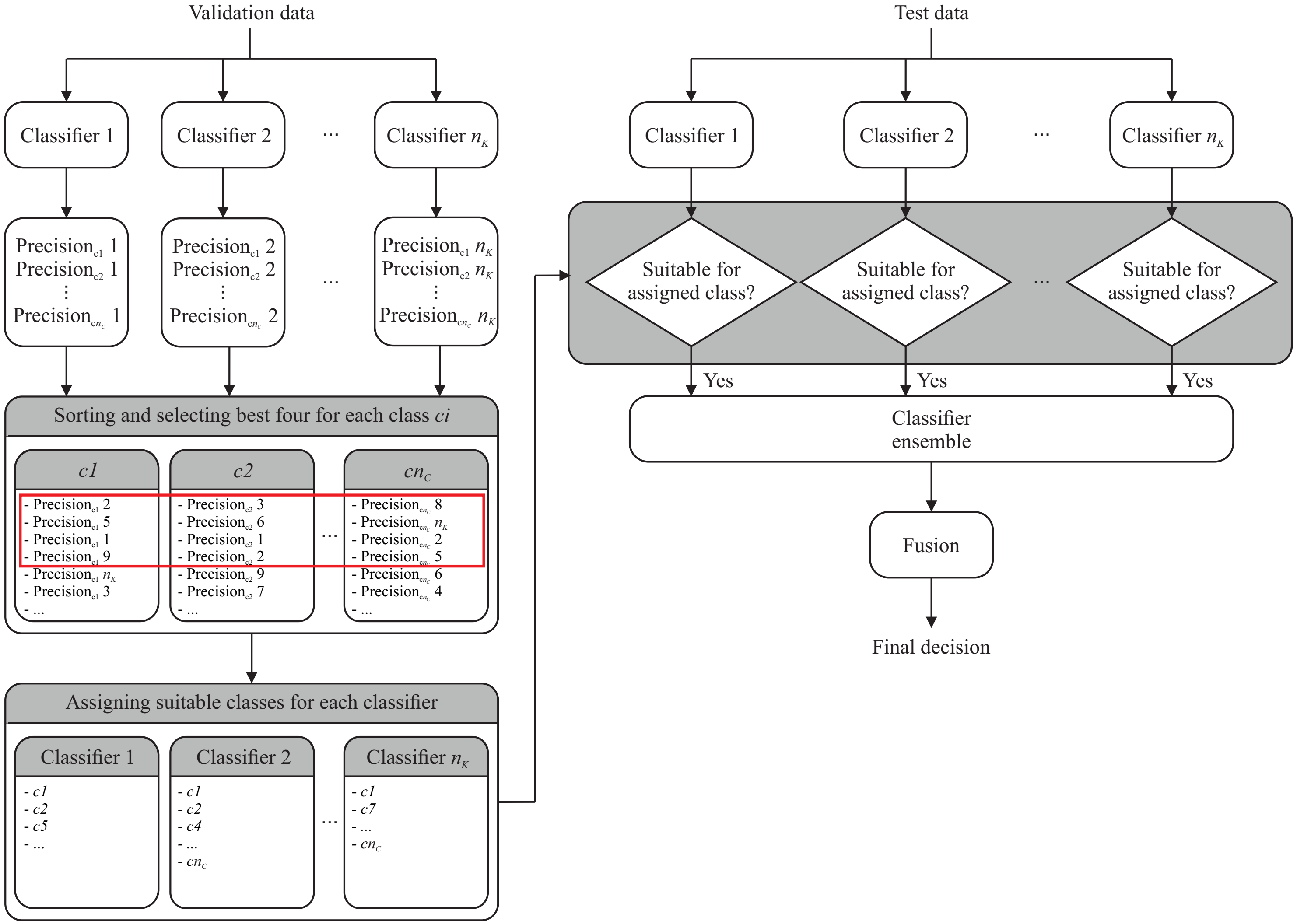

Figure 3.2: Concept of applied Dynamic Classifier Ensemble strategy.

During the test, the classification assignment of each classifier is compared to the set of classes the classifier is suitable for. If the actual assigned class is in the set, this classifier is selected for the ensemble. The resulting ensemble is fused afterward using different fusion methods.

3.2 Numerical analysis providing structural information

To compare the different fusion results, benchmark data from the UCR Time Series Classification Repository [Che$^+$14] are used. The selected data sets and their numbers of classes and samples are shown in Table 3.1. The data sets have to be classified to use classifier fusion. Using WEKA, 11 different classification approaches are applied with each data set. The chosen classifiers are Bayes Net (C1), Decision Table (C2), Hoeffing Tree (C3), k-nearest neighbor (C4), Multilayer Perception (C5), Naive Bayes (C6), partial C4.5 Decision Tree (C7), Random Forest (C8), Random Tree (C9), Reduced-Error Pruning Tree (C10), and Sequential Minimal Optimization (C11). In Table 3.2 the corresponding accuracies of the individual classifiers are presented for the whole data sets. The maximum value of each data set is marked bold, the minimum is underlined.

The outputs of the WEKA-classifiers are crisp values on abstract level and can be used to calculate the precision and the confusion matrix according to Eq. 2.17. For applying DSC not only the classes are considered, also all subsets of classes. To get the basic assignments of the power set $m : 2^{\Theta}$, the basic assignments of the individual classes are used, e.g. the basic assignment of the combination of class A and B is calculated adding the basic assignments of the individual classes $m(A)$ and $m(B)$. If the basic assignments of all possible combinations are determined, the values of the individual classifier are normalized to fulfill the requirement $\sum_{A \subseteq \Theta} m(A) = 1$.

Table 3.1: Applied benchmark data [RS16].

No.	Name	No. of classes	No. of samples
1	CBF	3	900
2	Coffee	2	28
3	ECG200	2	100
4	FaceFour	4	88
5	GunPoint	2	150
6	TwoPatterns	4	4000
7	Yoga	2	3000

Table 3.2: Accuracy in [%] of the individual classifiers for entire data sets no. 1-7 (maximum value in bold and minimum value underlined for each data set) (modified after [RS16]).

	1	2	3	4	5	6	7
C1	85.44	64.29	75.00	**89.77**	85.33	46.23	60.60
C2	76.78	64.29	70.00	54.55	90.67	43.45	59.50
C3	89.33	67.86	77.00	81.82	78.67	45.70	54.23
C4	85.00	75.00	**89.00**	87.50	92.00	**90.60**	**83.30**
C5	85.33	**96.43**	84.00	87.50	93.33	89.65	74.37
C6	**89.67**	67.86	77.00	84.09	78.67	45.68	54.23
C7	67.33	57.14	79.00	64.77	77.33	61.38	72.53
C8	87.11	60.71	81.00	**89.77**	**94.00**	83.18	80.53
C9	67.89	71.43	80.00	53.41	85.33	59.10	71.80
C10	68.22	64.29	74.00	44.32	80.67	56.53	65.83
C11	87.67	**96.43**	81.00	88.64	80.00	82.20	62.47

3.2.1 Results using Static Classifier Ensemble

To compare the different influences of the classifiers with high and low accuracy, eight different combinations (ensembles) of classifiers are chosen. These combinations are set based on the overall accuracy of the individual classifiers (Table 3.2) for each data set and contain

- the two classifiers with the highest accuracy (SCE A),
- the two classifiers with highest and lowest accuracy (SCE B),
- the two classifiers with medium accuracy (SCE C),
- the two classifiers with the lowest accuracy (SEC D),
- the five classifiers with the highest accuracy (SCE E),
- the two classifiers with highest, two with lowest and one with medium accuracy (SCE F),
- the five classifiers with medium accuracy (SCE G), and
- the five classifiers with the lowest accuracy (SEC H).

Table 3.3: Accuracy in [%] of the test data sets no. 1-7 for different SCE and fusion methods (modified after [RS16]).

No.		1	2	3	4	5	6	7
SCE A	WV	89.33	**96.43**	88.00	90.91	**95.33**	92.28	**83.30**
	BCR	89.44	**96.43**	88.00	93.18	**95.33**	92.28	**83.30**
	DSC	89.33	**96.43**	88.00	90.91	**95.33**	92.28	**83.30**
SCE B	WV	88.44	**96.43**	**89.00**	89.77	94.00	90.60	**83.30**
	BCR	89.67	**96.43**	**89.00**	92.05	94.00	90.70	**83.30**
	DSC	88.44	**96.43**	**89.00**	89.77	94.00	90.60	**83.30**
SCE C	WV	89.89	67.86	80.00	84.09	84.67	61.85	65.83
	BCR	**90.44**	67.86	80.00	84.09	84.67	62.63	65.83
	DSC	89.89	67.86	80.00	84.09	84.67	62.88	65.83
SCE D	WV	73.89	64.29	76.00	60.23	78.67	44.03	54.23
	BCR	74.56	64.29	76.00	65.91	78.67	46.05	54.23
	DSC	73.89	64.29	76.00	62.50	78.67	44.58	54.23
SCE E	WV	89.78	92.86	85.00	90.91	94.67	92.88	81.73
	BCR	90.00	**96.43**	85.00	94.32	94.67	**93.93**	81.93
	DSC	89.78	**96.43**	85.00	90.91	94.67	92.98	83.73
SCE F	WV	88.22	**96.43**	83.00	95.45	91.33	90.70	72.93
	BCR	89.89	**96.43**	86.00	**97.73**	94.00	92.20	82.37
	DSC	89.78	**96.43**	86.00	95.45	94.00	91.70	82.37
SCE G	WV	87.22	67.86	80.00	89.77	90.00	76.90	69.63
	BCR	**90.44**	67.86	78.00	**97.73**	91.33	81.68	73.73
	DSC	90.33	67.86	78.00	92.05	90.67	80.90	73.73
SCE H	WV	83.44	64.29	73.00	80.68	82.67	49.90	62.97
	BCR	88.22	60.71	71.00	93.18	84.67	54.13	60.87
	DSC	88.22	60.71	71.00	88.64	83.33	53.93	60.87
Mean_i		80.88	71.42	78.81	75.10	85.09	63.97	67.21
Max_i		89.67	**96.43**	**89.00**	89.77	94.00	90.60	**83.30**

The accuracies of the fused results are shown in Table 3.3 for the different ensembles and fusion methods. The first four ensembles (SCE A to SCE D) are used to describe the combination of two, the last four ensembles (SCE E to SCE H) the combination of five classifiers. The mean value of the accuracies of the individual classifiers ($\mathrm{Mean}_{\mathbf{i}}$) as well as the best individual classifier accuracy ($\mathrm{Max}_{\mathbf{i}}$, compare Table 3.2 (bold)) are also shown in Table 3.3. The results with higher accuracy than the maximum accuracy of the individually used classifiers are marked with dark green cells. If the accuracy is less than the maximum individual accuracy, but greater than $\mathrm{Mean}_{\mathbf{i}}$, the cell is highlighted in light green. An accuracy less than $\mathrm{Mean}_{\mathbf{i}}$ is marked in light red. The best fusion result (highest accuracy) of each data set is highlighted in bold. None of the fused accuracies is less than the lowest accuracy of the individual classifiers (compare Table 3.2 (underlined)).

Comparing all accuracies, for each data set the accuracy of combined classifiers is greater or equal than the maximum accuracy of individual classifiers. For four data sets (1, 4, 5, and 6), the overall accuracy is improved compared to the individual accuracy (see dark green color). The accuracy for most data sets can be improved using the five best classifiers (SCE E). Compared to the combination of two classifiers (Table 3.3 (top)), the combination of five classifiers (Table 3.3 (bottom)) show better results, e.g. more dark green cells using SCE E as using SCE A, and less red cells using SCE H as using SCE D. As long as the best classifiers are in the set, the accuracy is always greater than $\mathrm{Mean}_{\mathbf{i}}$ (SCE A, SCE B, SCE E, and SCE F).

Comparing the fusion methods, there is just a small difference between the results. Comparing the cells with dark green color (improvement of accuracy), it can be stated that BCR can achieve the most improvement, followed by DSC and WV. Using different SCE with BCR results in the best accuracy for each data set ((Table 3.3 (bold)). From this numerical comparison questions regarding the best static ensemble in combination with the best fusion method can be answered briefly: The combination of SCE E with BCR will realize the best results in the case the data are structured similar as those used for this numerical study based on completely independent data sets.

Lessons learned From the numerical analysis using SCE out of eleven classifiers, it can be concluded, that using five classifiers in an ensemble shows better performance than using only two. Using fusion, the selection of suitable classifiers out of the total set of available classifiers is recommended to ensure the same or even higher accuracy than the best individual classifier. Not for all data sets the combination of classifiers with the best individual accuracy lead to the best fusion performance. Considering SCE, BCR should be preferred over DSC and WV.

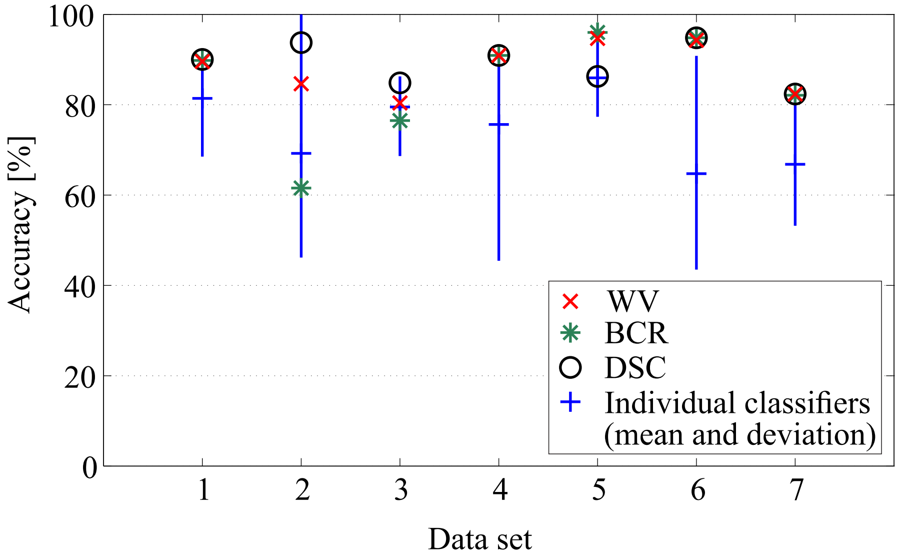

Figure 3.3: Accuracy of the individual and combined classifiers for the test data of all data sets (no. 1-7) using DCE [RS16].

3.2.2 Results using Dynamic Classifier Ensemble

First, one half of data is assigned as validation and the other half as test data. The results of the fusion results using DCE are shown in Figure 3.3. The blue lines represent the range from the minimum to the maximum accuracy of the individual classifiers.

Compared to the results in Chapter 3.2.1, the BCR show worse results for the data sets Coffee (no. 2) and ECG200 (no. 3). The accuracy of the data sets CBF (no. 1), FaceFour (no. 4), TwoPatterns (no. 6), and Yoga (no. 7) is nearly the same for the three different fusion methods.

To ensure, that the results shown in Figure 3.3 are independent from validation and test data, six variations of the validation and test data sets are considered. Two data sets with most similar results of the fused accuracy and two with the less similar results are chosen. For the data sets CBF (no. 1), Coffee (no. 2), ECG200 (no. 3), and Yoga (no. 7) the validation and test data sets are partitioned using 3-fold cross-validation. The validation and test data sets are also exchanged, so two third of data for validation and one third for test are also used.

In Figure 3.4 the range of accuracy from maximum to minimum value as well as the mean value for individual classifiers and fused results are shown. The data sets with similar results in Figure 3.3 (no. 1 and no. 7) also show similar results using cross-validation. Also the data sets with different results in Figure 3.4 (no. 2 and no. 3) have a similar mean value if cross-validation is used, but the deviation of results is

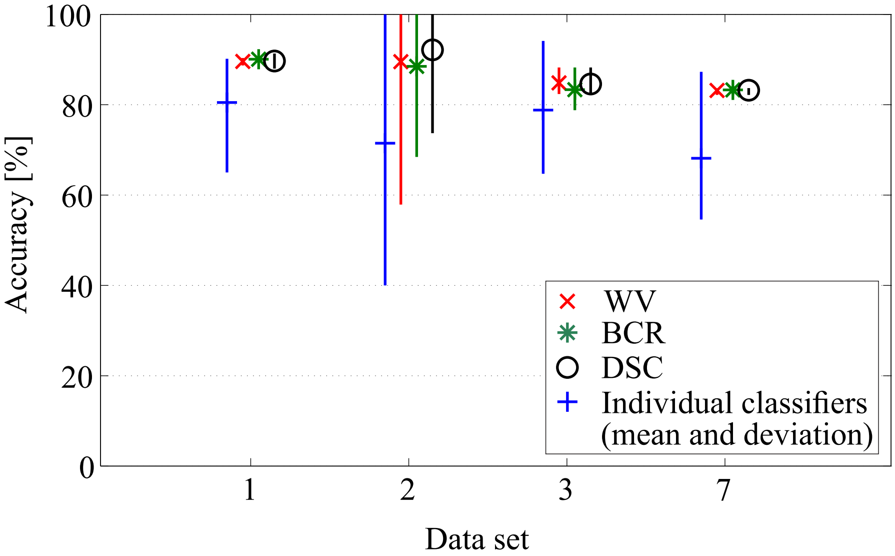

Figure 3.4: Accuracy of the individual and combined classifiers for the test data set using cross-validation and DCE [RS16].

greater. Probably, the greater deviation is caused by the smaller number of samples available for validation and test (see Table 3.1). The fusion method DSC has least deviation compared to the other fusion methods, which could lead to the conclusion, that for this numerical example the DSC method is most robust against validation data set change. From the results using DSC the question regarding the best fusion method can be answered: The DSC will realize the best results independent from the randomly compiled validation and test data.

Lessons learned Considering DCE, the best results regarding the overall accuracy as well as the independence to changes of the validation and test data set can be achieved using DSC.

3.3 Application to experimental data

For further analysis using a smaller set of available classifiers and application data for condition monitoring, two experimental examples are used. The SCE and DCE strategies explained are considered to show the influence of the selection strategy to the overall performance.

3.3.1 Fault diagnosis of hot rolling mills

The introduced ensemble strategies are applied to real industrial data from hot rolling mills (HRM). During rolling process in the field of metal industry the geometry, surface condition, and thickness of the strip, plate, or sheet must meet given requirements from the customer. Deviations in strip travel lead to influences on product quality and cause down times. As illustrated in [RJS14], the events can be classified as

- Stable rolling with strip in stand (Class 1),
- Stable rolling without strip in stand (Class 2),
- Fault cobble (Class 3), and
- Fault shearing tale (Class 4).

In [RJS15] five different time-frequency-based analysis methods are applied to the detection of cobble as a fault in rolling process: continuous and discrete wavelet transform (CWT/DWT), empirical mode decomposition (EMD), short-time Fourier transform (STFT), and Wigner-Ville distribution (WVD). As introduced in [RJS14], 80 samples (20 per class) are filtered using the mentioned methods and classified using two different classification approaches: support vector machine (SVM) and cross-correlation (CC). This leads to six different filter-classifier combinations (approaches): CWT-SVM, DWT-SVM, EMD-CC, EMD-SVM, STFT-SVM, and WVD-SVM. The individual classification results as well as the mean value of the individual accuracy are stated in Table 3.4.

Again the best result is marked in bold, the worst result is underlined. The performance of the individual classifiers strongly vary from 30.00 % to 85.00 %, where

Table 3.4: Individual filter-classifier combination accuracy.

No.	Approach	Accuracy [%]
1	CWT-SVM	65.00
2	DWT-SVM	**85.00**
3	EMD-CC	82.50
4	EMD-SVM	72.50
5	STFT-SVM	<u>30.00</u>
6	WVD-SVM	72.50
Mean		67.91

the classifiers no. 4 (EMD-SVM) and no. 6 (WVD-SVM) have the same individual accuracy. Using six different classifiers allows to show the fusion results using all possible ensembles generated from the set of these six classifiers.

In Table 3.5 the accuracy for all SCEs and DCE using the three fusion methods WV, BCR, and DSC are shown. The numbers denote the combined classifiers (e.g. 1,2 means the fusion of the results classified by no. 1 (CWT-SVM) and no. 2 (DWT-SVM) (see Table 3.5)). The color notation is the same as in Chapter 3.2.1.

Considering the combination of two classifiers, 15 different combinations of the individual classifiers are possible. The fusion of the two worst classifiers (1,5) leads to the worst results of all ensembles (underlined) and a deterioration compared to the mean performance of the individual classifiers (red color). Combining the two best classifiers (2,3), the fused results do not show best performance, not even an improvement compared to the best individual. Best results combining two of the available six classifiers can be achieved using BCR or DSC for fusion of no. 2 and no. 6 or using DSC for fusion of no. 2 and no. 4. Using BCR, the ensemble 2,6 lead to the best performance compared to all other ensembles. Although classifiers no. 4 and no. 6 show the same individual accuracy, the fusion results differ whether no. 4 or no. 6 is part of the ensemble.

This can also be observed using fusion of three individual classifiers. Fusion of two best individual ones with classifier no. 4 (2,3,4) generates better fusion accuracy than the fusion using no. 6 (2,3,6) regarding WV and DSC. Considering BCR the accuracy for the ensemble 2,3,4 is worse than the results fusing the ensemble 2,3,6. Considering the fusion of three worst classifiers, again the results differ depending on the used third classifier. Using WV and DSC, the ensemble 1,4,5 show worse results that fusing the ensemble 1,5,6, using BCR this is inversely. Regarding fusion of two, three, or four worst classifiers, BCR produces better results, than using WV and DSC, considering the combination of the best classifiers, BCR is always worse than WV and DSC independent from the number of fused classifiers. Combining three classifiers, the performance using WV or DSC and the ensemble 2,3,4 is the best.

Different to the fusion of three classifiers, the fusion of four best classifiers using WV do not result in the pest performance. In this case, the fusion of the best three (2,3,4) and the worst individual classifier (no. 5) produces the best results for all ensembles independent from the number of considered classifiers. Using DSC, the ensembles 2,3,4,5 and 2,3,4,6 result in the same fused accuracy, although classifier no. 4 has a higher individual accuracy (72.50 %) than classifier no. 5 (30.00 %).

Regarding the fusion of five classifiers, the ensemble 2,3,4,5,6 (neglecting classifier no. 1) generates the highest fusion accuracy using WV and DSC. Using DSC produces the same results for the ensemble 2,3,4,5,6 as for 2,3,4,5 and 2,3,4,6, which show the pest performance (97.50 %) in the complete considerations.

The fusion of all classifiers can not outperform the best individual classifier.

Using the DCE, the fused accuracy is 90.00 %, which is higher than the best individual, but is not outperforming the best SCE.

Table 3.5: Accuracy in [%] using all possible ensemble selection strategies (part 1).

Ensemble selection strategy		WV	BCR	DSC
SCE	1,2	80.00	82.50	85.00
	1,3	77.50	75.00	77.50
	1,4	70.00	75.00	70.00
	1,5 (2 worst)	65.00	67.50	65.00
	1,6	72.50	70.00	70.00
	2,3 (2 best)	85.00	77.50	85.00
	2,4	85.00	85.00	87.50
	2,5	85.00	82.50	82.50
	2,6	82.50	87.50	87.50
	3,4	80.00	77.50	80.00
	3,5	82.50	80.00	82.50
	3,6	77.50	77.50	77.50
	4,5	72.50	72.50	72.50
	4,6	75.00	72.50	70.00
	5,6	77.50	80.00	80.00
	1,2,3	87.50	77.50	87.50
	1,2,4	80.00	85.00	80.00
	1,2,5	75.00	82.50	87.50
	1,2,6	75.00	82.50	75.00
	1,3,4	85.00	77.50	82.50
	1,3,5	77.50	72.50	77.50
	1,3,6	75.00	72.50	75.00
	1,4,5 (3 worst)	67.50	77.50	67.50
	1,4,6	75.00	75.00	72.50
	1,5,6 (3 worst)	72.50	75.00	72.50
	2,3,4 (3 best)	95.00	77.50	95.00
	2,3,5	85.00	77.50	87.50
	2,3,6 (3 best)	92.50	80.00	92.50
	2,4,5	77.50	80.00	82.50
	2,4,6	82.50	85.00	82.50
	2,5,6	80.00	85.00	87.50
	3,4,5	77.50	75.00	80.00
	3,4,6	90.00	77,50	87.50
	3,5,6	77.50	75.00	77.50

Table 3.5: Accuracy in [%] using all possible ensemble selection strategies (part 2).

Ensemble selection strategy		WV	BCR	DSC
	4,5,6	72.50	80.00	77.50
	1,2,3,4	87.50	77.50	92.50
	1,2,3,5	87.50	75.00	87.50
	1,2,3,6	82.50	77.50	82.50
	1,2,4,5	80.00	82.50	80.00
	1,2,4,6	75.00	82.50	75.00
	1,2,5,6	75.00	82.50	80.00
	1,3,4,5	85.00	75.00	80.00
	1,3,4,6	82.50	77.50	82.50
	1,3,5,6	75.00	72.50	80.00
	1,4,5,6 (4 worst)	75.00	77.50	75.00
	2,3,4,5	**97.50**	75.00	**97.50**
	2,3,4,6 (4 best)	92.50	77.50	**97.50**
	2,3,5,6	92.50	77.50	92.50
	2,4,5,6	85.00	82.50	82.50
	3,4,5,6	87.50	75.00	87.50
	1,2,3,4,5	87.50	75.00	92.50
	1,2,3,4,6 (5 best)	85.00	77.50	85.00
	1,2,3,5,6	85.00	75.00	85.00
	1,2,4,5,6	75.00	82.50	77.50
	1,3,4,5,6 (5 worst)	82.50	75.00	82.50
	2,3,4,5,6	95.00	75.00	**97.50**
	1,2,3,4,5,6 (all)	85.00	75.00	85.00
DCE		90.00	80.00	90.00

Lessons learned For the considered data set, the SCE strategy can lead to better performance than using DCE. Using SCE, the best results can be achieved using WV and the ensemble 2,3,4,5 as well as using DSC and the ensembles 2,3,4,5 (three best and the worst individual classifier), 2,3,5,6 (four best individual classifiers), or 2,3,4,5,6 (neglecting the second worst individual classifier). In summary, the combination of the best classifiers regarding the accuracy do not automatically lead to the best performance. Classifiers with the same accuracy can lead to different fusion performance.

3.3.2 Damage detection in composites

Composite materials are increasingly used to replace conventional construction materials due to advantageous mechanical properties, such as high specific strength resulting from their complex structure. However, the more widespread use is currently restricted by the lack of ductility compared to metallic materials as well as several micromechanical damage modes [PGO14]. Different NDT techniques, which indicate developing damages, are proposed for continuous monitoring of these materials to ensure equal degree of reliability and safety [Cri$^+$15]. In [WBS18], experiments were conducted using a bending test rig to subject specimens of composite material to cyclic bending load with different amplitudes and frequencies. Data samples with known labels are necessary to validate and test the classifier and fusion performance. Therefore 120 data samples are selected and manually classified. To assess the reliability automatic damage detection techniques, a showcase algorithm for damage classification in composites has been implemented using Acoustic Emission measurements and Support Vector Machine (SVM). To combine the results, also the K-Nearest Neighbor (KNN) and Artificial Neural Network (ANN) methods are used to classify the same data sets. Using the data sets with known labels, results from a Reproducing Kernel Hilbert Space (RKHS) method are also considered. The method is based on the ideas given in [KN14], but uses RKHS with Gaussian kernels which can be seen as a Radial Basis Function Network with Gaussian basis functions placed on every data point. Caused during experiments, four different damages are considered: delamination (class 1), matrix crack (class 2), debonding (class 3), and fiber breakage (class 4). Class 5 is denoted as noise.

The performance measures of the individual classification results of ANN, KNN, RKHS, and SVM are given in Table 3.6. As performance measures the accuracy, precision, recall (detection rate), and the false alarm rate (FAR) are used to evaluate

Table 3.6: Performance measures of individual classifiers and fused results using same data for validation and test (modified after [Rot$^+$17]).

No.	Approach	Accuracy [%]	Precision [%]	Recall [%]	FAR [%]
1	ANN	92.50	92.52	92.50	1.87
2	KNN	95.00	95.13	95.00	1.25
3	RKHS	**96.67**	96.66	**96.67**	**0.83**
4	SVM	**96.67**	**96.73**	**96.67**	**0.83**
Mean		95.21	95.26	95.21	1.20
WV		95.83	95.87	95.83	1.04
BCR		97.50	97.60	97.50	0.62

the classification and fusion results. The best value in one column is marked with bold letters. Regarding the accuracy, recall, and FAR, the classifiers RKHS and SVM are the best out of the four considered ones. Only the precision value is slightly higher for SVM than using RKHS. Using WV as fusion method, the performance can not be improved. The fusion using BCR shows higher performance measure values for all measures.

Using different variations of validation and test data generated from the total data set, the results differ from the results in Table 3.6. The performance measures of six different variations of validation and test data sets are shown in Table 3.7. Both, WV and BCR show worse performance as the individual classifiers RKHS and SVM, although the accuracy, recall and FAR of the individual classifiers are same as using only one validation and test data set distribution and precision value has just slightly changed. To improve the fusion performance, a Static Ensemble Selection is used. In Table 3.8 the accuracy using WV and BCR in combination with all possible SCEs and the use of DCE are presented. For this data set, the fusion of the first (ANN) and the fourth (SVM) classifier shows the best fusion results, even though the classifier ANN has the worst individual classification performance. Fusion performance of BCR is less than training with all data, although the results of individual are just slightly changing. This leads to the conclusion, that the fusion performance is not only depending on the performance of the individual classifiers, but also on the data itself.

Lessons learned The results show that the performance can be improved using fusion methods or even ensemble selection strategies, but the performance depends on the used fusion method, the validation data set, and also on the selected classifiers. In the given example, the fusion using BCR with the performance measures

Table 3.7: Performance measures of individual classifiers and fused results using variation of validation and test data sets (modified after [Rot+17]).

No.	Approach	Accuracy [%]	Precision [%]	Recall [%]	FAR [%]
1	ANN	92.50	92.60	92.50	1.87
2	KNN	95.00	95.29	95.00	1.25
3	RKHS	**96.67**	96.70	**96.67**	**0.83**
4	SVM	**96.67**	**96.78**	**96.67**	**0.83**
Mean		95.21	95.34	95.21	1.20
WV		95.83	95.98	95.83	1.04
BCR		95.83	95.98	95.83	1.04

Table 3.8: Accuracy in [%] using ensemble selection strategies and variation of validation and test data sets.

Ensemble selection strategy		WV	BCR
SCE	1,2 (2 worst)	93.06	94.17
	1,3	94.45	95.00
	1,4	96.11	**96.67**
	2,3	94.45	95.00
	2,4	94.72	95.56
	3,4 (2 best)	95.83	96.39
	1,2,3	94.72	94.72
	1,2,4	95.28	95.28
	1,3,4	95.83	95.83
	2,3,4 (3 best)	95.83	95.83
	1,2,3,4 (all)	95.83	95.83
DCE		94.72	95.28

from the complete data set shows the best results. In case of variation of validation and test data, the selection of ANN and SVM to be fused has the most improvement of performance.

3.4 Conclusions from the comparison using ensemble selection

In the previous chapters, the question regarding the best selection strategy is analyzed. Therefore the first question is, whether the SCE or DCE should be preferred. From the results it can be stated, that both, the applied SCE and DCE can improve the overall performance. In this analysis the SCE is outperforming the DCE in most cases.

The question regarding the optimal number of classifiers in an ensemble is depending on the total number of available classifiers. Selecting an ensemble out of eleven base classifiers, the ensemble should rather have five instead of two classifiers. Using six base classifiers (as in the experimental example of fault diagnosis of hot rolling mills), the most improvement can be achieved using four classifiers. The experimental results of damage detection in composites, where four base classifiers are available, the best result is achieved using two classifiers. An optimal number of classifiers can not generally be given, but in all cases, the number is around half of the base classifiers. In all cases, the fusion of all available base classifiers leads not to the best fusion performance.

Considering the question of the properties a classifier should have in an ensemble to improve the results, the answer is contrary to the expectation, that the combination of the best classifiers always lead to the best results. The results of the numerical and experimental analysis show, that also the combination of best classifiers with one or two worse classifiers can lead to better performance than combining the best individual classifiers. A change of used training and test data sets without significant change in the performance measures show different fusion results. This results raise the question whether a change of the individual performance of one of the selected classifiers is influencing the overall performance. This question is analyzed in the next chapter. The influence of the data characteristic on the fusion performance is considered in Chapter 6.

A conclusion about the best fusion method can not be answered yet, because for different data sets different fusion methods are outperforming the others. In most cases the BCR method is slightly better than WV and DSC, but in case of the fault diagnosis of hot rolling mills the BCR is worse than WV and DSC.

4 Analysis of classifier properties influencing fusion performance

To define requirements for a good fusion performance and to evaluate the potential for higher accuracy, in this chapter the idea of a fictional classifier is introduced. The idea introduced here is to change precision values for one of the classifiers in order to evaluate options to improve overall fusion accuracy. Using this fictional classifier, the influence of the precision values on the overall fusion performance is visible. The goal is to get the optimal precision values to see, if they are generalizable for different ensembles or data sets. From the results, some important challenges can be solved: Can the performance of one individual classifier improve the overall accuracy? Can the performance of the fused results be improved by changing performance measures of the fictional classifier? Which properties should an additional classifier have to improve or not deteriorate the fusion performance? The considerations of varying precision values also allows the establishment of a supervised strategy to adapt precision values in order to get better fusion results for unknown samples. For illustrating the effects a practical example based on four benchmark data sets is used to analyze the influences. The introduced methods are applied to the real example of fault diagnosis of hot rolling mills. The results show that using a fictional classifier the overall accuracy can be outperformed depending on data sets.

In Chapter 4.1 the new idea of the fictional classifier and the training-based fusion algorithm are presented. Chapter 4.2 shows the results of numerical analysis using four benchmark data sets with distinction between two classes. The experimental results distinguishing between four classes are discussed in Chapter 4.3. In Chapter 4.4 conclusions from the results are given.

Most of the content, figures, and tables in this chapter are prepared for publication of [RS19].

4.1 Concept of fictional classifier

To analyze the requirements for an improvement of reliability using decision fusion, the concept of a fictional classifier (FC) is introduced based on the idea that the precision value has a significant influence on the fusion performance. In case of using BCR for a two class problem, the precision value is the value used for calculating the belief values. Therefore the influence on fused accuracy by changing the precision values should be analyzed. Validation data are used to show the possible improvements of the fused accuracy. The scheme of the concept is shown in Figure 4.1. All classifiers of the static classifier ensemble classify the validation data, but one is set as fictional classifier. Classification assignments stay the same, but precision values, which are used to calculate the belief value, are now considered as variable.

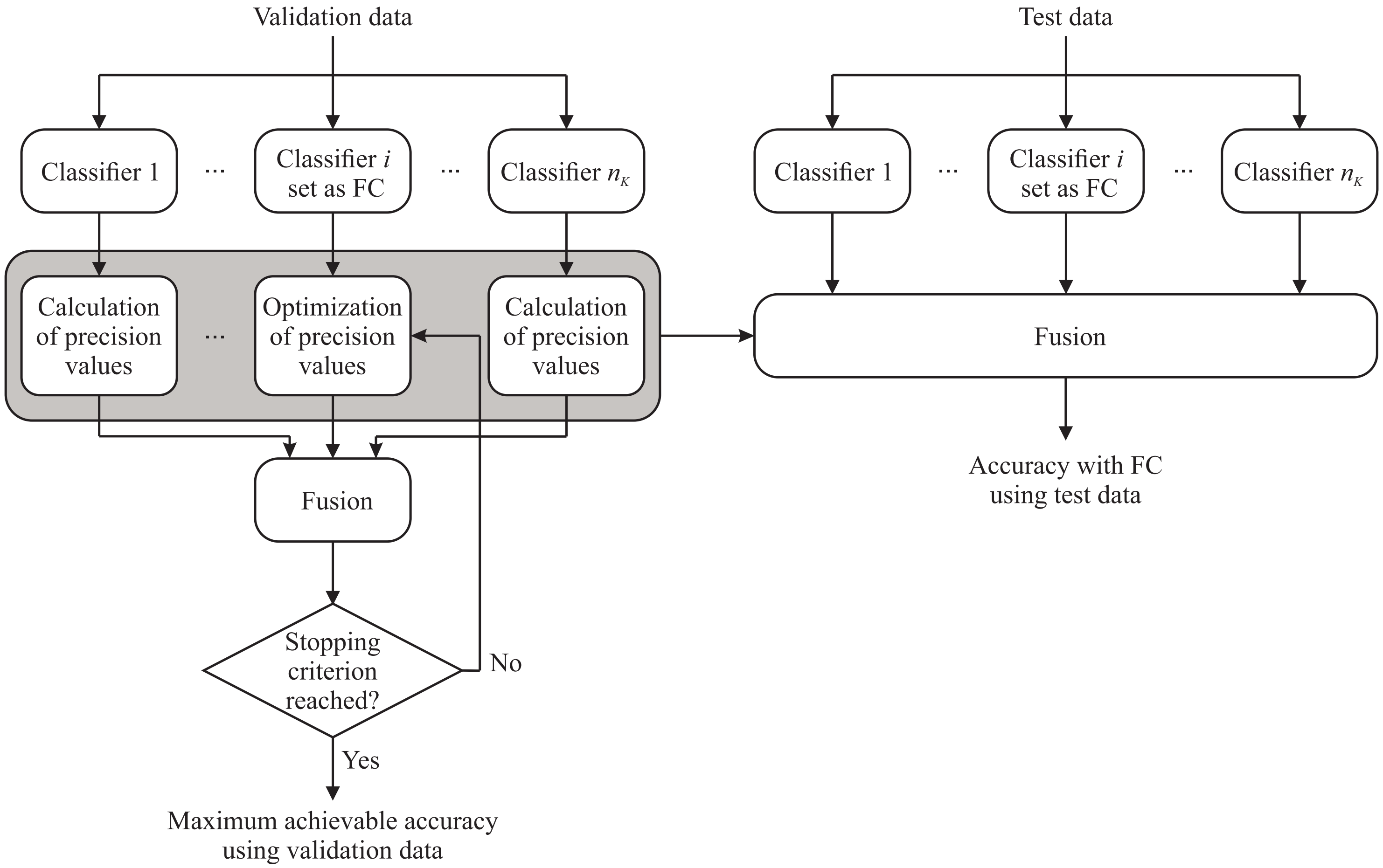
Validation data
Classifier 1
...
Classifier *i* set as FC
...
Classifier n_K
Calculation of precision values
...
Optimization of precision values
Calculation of precision values
Fusion
Stopping criterion reached?
No
Yes
Maximum achievable accuracy using validation data
Test data
Classifier 1
...
Classifier *i* set as FC
...
Classifier n_K
Fusion
Accuracy with FC using test data

Figure 4.1: Concept of fictional classifier (FC) [RS19].

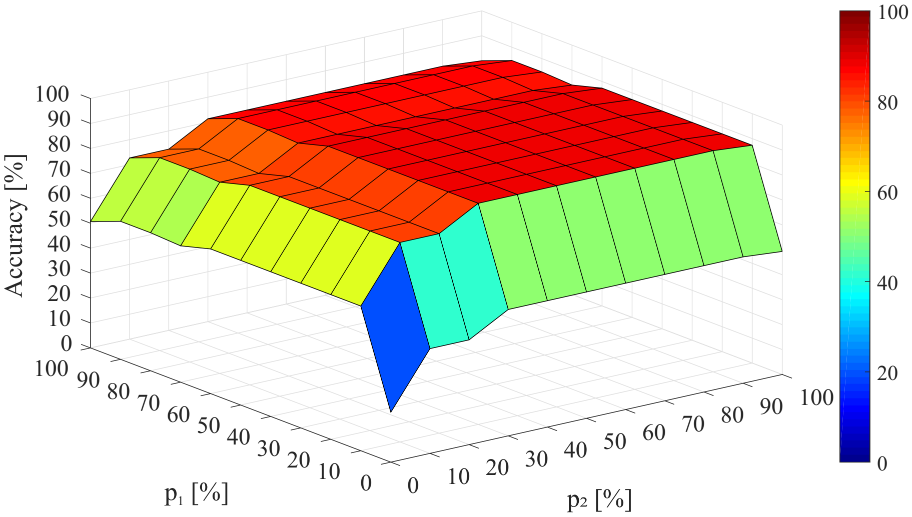

Figure 4.2: Example describing the dependency of accuracy for different precision values p_1 and p_2 [RS19].

In the case of two considered classes the probability matrix of FC will be

$$P^{FC} = \begin{bmatrix} p_1 & 1-p_2 \\ 1-p_1 & p_2 \end{bmatrix}, \tag{4.1}$$

with p_1 and p_2 as variable precision values of class 1 and class 2 respectively. Using different values for p_1 and p_2, the influence of the precision values of one classifier on the overall accuracy of fused results can be shown. As an example for changing accuracy by varying precision values, in Figure 4.2 the accuracy of fused results depending on the precision values p_1 and p_2 is shown. The fused accuracy is calculated based on the use of different values for p_1 and p_2 in the range of $[0\,\%,\ 100\,\%]$. Only for a specific range of p_1 and p_2 the maximum accuracy of fused results can be reached. For this example values in the intervals $p_1 = [0\ \%,\ 50\ \%]$ and $p_2 = [30\ \%,\ 90\ \%]$ lead to the maximum accuracy of 89 %.

Using non-dominated sorting genetic algorithm II (NSGA-II) [Son14], the values of p_1 and p_2 as decision variables leading to the maximum reachable fused accuracy (objective) can be determined. In each step, the decision variables are changed according to an evolutionary algorithm. Depending on the resulting accuracy, the new generation is generated. In most cases, especially for considering two classes, there is more than one precision value set leading to the same fused accuracy, as for the example in Figure 4.2.

Based on the concept of fictional classifier, the idea is extended to a training-based fusion algorithm (see Figure 4.1). In training, the precision values of the FC are optimized and selected by calculating the mean value and finding the nearest neighbor. Using this parameter set, unknown data samples (test data) can be fused.

4.2 Numerical analysis

Benchmark data from the UCR Time Series Classification Repository [Che$^+$14] are used to analyze the dependencies between precision values of FC and accuracy of fused results. Therefor four data sets with two classes are chosen: Coffee (28 samples), ECG200 (100 samples), GunPoint (150 samples), and Yoga (3000 samples). The data sets have to be classified to use classifier fusion. Using WEKA [Hal$^+$09], the same 11 different classification approaches are applied with each data set as in Chapter 3.2. The results of the individual classification are also presented in Chapter 3.2. For the four data sets, different individual classifiers are better or worse, so that no classifier can be denoted as best. The results analyzed in Chapter 3.2 show, that only for data set GunPoint the best individual classifier can be outperformed using fusion of a selected ensemble. Therefore a combination of this classifiers using fusion methods as well as using an FC can be realized.

4.2.1 Classifier included in ensemble set as fictional classifier

Different static classifier ensembles are used to examine the results for various combinations of classifiers with different performances. The selection of ensembles is based on the accuracy of individual classifiers. The nine ensembles and their abbreviations are shown in Tab. 4.1. In each ensemble, each classifier is set as FC consecutively. The data sets are divided into three training and test data sets using 3-fold cross-validation.

The simulation results for the data sets are shown in Figure 4.3 - 4.6. In the first row the results from training and in the second row from test are shown. The figure on top left shows the suitable parameter ranges achieving maximum accuracy of fused results for each classifier ensemble with varying classifier set as FC. The corresponding maximum accuracy achievable as well as the maximum individual accuracy of

Abbreviation	Ensemble selection
2B/5B	Two/five best classifiers
2W/5W	Two/five worst classifiers
1B+1W/2B+2W	One/two best and one/two worst classifiers
3M/5M	Three/five medium-good classifiers
ALL	All eleven classifiers

Table 4.1: Static classifier ensembles selected by accuracy of individual classification results applied on benchmark data [RS19].

the ensemble and the accuracy of fused results using the initially calculated precision values are shown in the figure top right. On bottom left the selected parameter sets are shown, which lead to an accuracy of fused results (shown in figure on bottom right) combining FC with selected parameters and the other classifiers of this ensemble.

The results of training phase for data set Coffee shows that the range of precision values leading to the maximum achievable accuracy is different for different classifier ensembles (see Figure 4.3 top left). In the case of 2B, 5B, and ALL the range for p_1 and p_2 varies between 0 % and 100 %. This means, for all parameter sets in between this values, the accuracy is the same. In case of 2W and 5W (combining the worst classifiers) the range for suitable precision values is smaller and the resulting accuracy is higher than the accuracy achievable without FC and the maximum individual accuracy (top right). Not always a smaller range of precision values lead to a higher achievable accuracy (5M, 1B+1W, and 3M). The specific value ranges of p_1 and p_2 also varies for different classifiers set as FC. For example in case of combining one best and one worst classifier (1B+1W), the range is smaller for the second classifier (best classifier) set as FC. The selected values both for p_1 and p_2 are nearly the same and vary around 50-60 % (bottom left). Combining the five worst classifiers (5W) the selected parameters show the most deviation for different data sets. Also the deviation of accuracy (bottom right) leads to the dependency of results from the considered training and test data sets. The maximum achievable accuracy in training phase can not be reached in test. In all cases the fusion with FC can not outperform the best individual classifier, but compared to the fusion without FC the accuracy is the same or higher.

The simulation results of data set ECG200 (Figure 4.4) lead to similar conclusions, although there is a higher dependency of precision value ranges and achievable accuracy on different classifiers set as FC (top left and top right). In Figure 4.4 it can be seen, that the selected parameter values (bottom left) vary more significantly compared to data set Coffee, especially by fusing the ensembles 5M and ALL. Otherwise the deviation of accuracy (bottom right) is less. Again the highest achievable accuracy from training can not be reached in the test. In most cases the accuracy is less than the best individual and also sometimes less than the fusion without FC.

The improvement of achievable accuracy using the data set GunPoint is higher than for the other data sets (see Figure 4.5 top right). In most cases the accuracy of fused results using FC is higher than the best individual fusion accuracy. Also the range of precision values leading to this higher accuracy is lower than for the data sets Coffee and ECG200 (top left). From the figure bottom left it can be seen that the selected values of p_1 and p_2 differ from each other. This can not be clearly detected from the other data sets, which leads to the conclusion, that for this data set the classifiers show different behaviors for each class and a different precision value for each class leads to better results in the accuracy. Using the selected parameter sets for the FC, the accuracy of the fused results can be improved in some cases (3M, 5W, 5M), but can not outperform the individual best like in training.

For data set Yoga a very specific range of precision values is resulting in the maximum achievable accuracy (Figure 4.6 top left), but the absolute values are varying for the different ensembles. A higher accuracy than the accuracy of fusion without FC can be obtained, but again the best individual performance can not be improved (top right). Nevertheless, the results from training can be achieved also in test using the selected parameter (bottom left) and again the fusion results using an FC are better than the results of using normal BCR. Additionally the deviation of results is very small. This data set contains more samples, than the others, which can be the reason for different results. To proof this, more data sets with high and low number of samples should be analyzed.

Lessons learned Summarizing the analysis using a fictional classifier as one classifier from a selected ensemble based on benchmark data sets, some conclusions can be drawn to generalize the (numerical) results:

- Resuming all example combinations it can be concluded that in general a higher accuracy can be obtained using the FC compared to the accuracy of fusion without FC. The best individual performance can not be outperformed.
- The best individual performance for one data set does not always result from the same classifier, so that in case of data changes, which are not trained, the fusion performance using an FC is more reliable than without using FC.
- From the results it can be detected that no common value of p_1 and p_2 can be concluded for all ensembles or data sets. The precision values are very different for specific ensembles, data sets, and in some cases also for different classifiers set as FC.
- Using the in training selected precision values to new data, fusion results can be improved.
- In case of a higher data sample number (example: Yoga), the results are more specific, the training and test results are nearly the same, but for lower number of data samples (example: Coffee) the variation of precision values as well as achievable accuracy is very high.
- Using varying precision values, the data sets with improvement potential using ensembles and classifiers set as FC can be stated.
- Combining classifiers with low accuracy, the improvement potential affected by the precision values is higher.

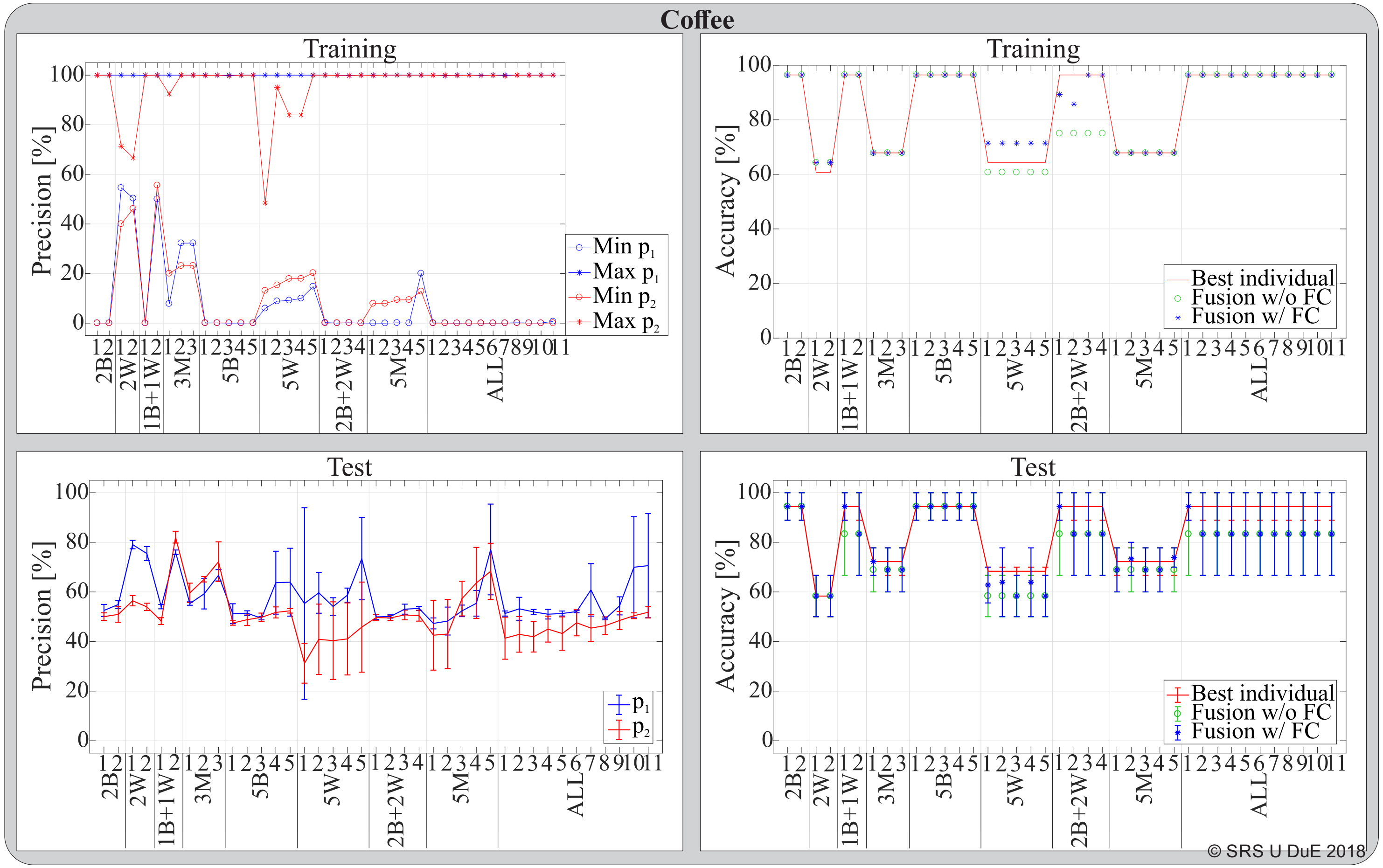

Figure 4.3: Simulation results of data set Coffee (FC included in ensemble) [RS19].

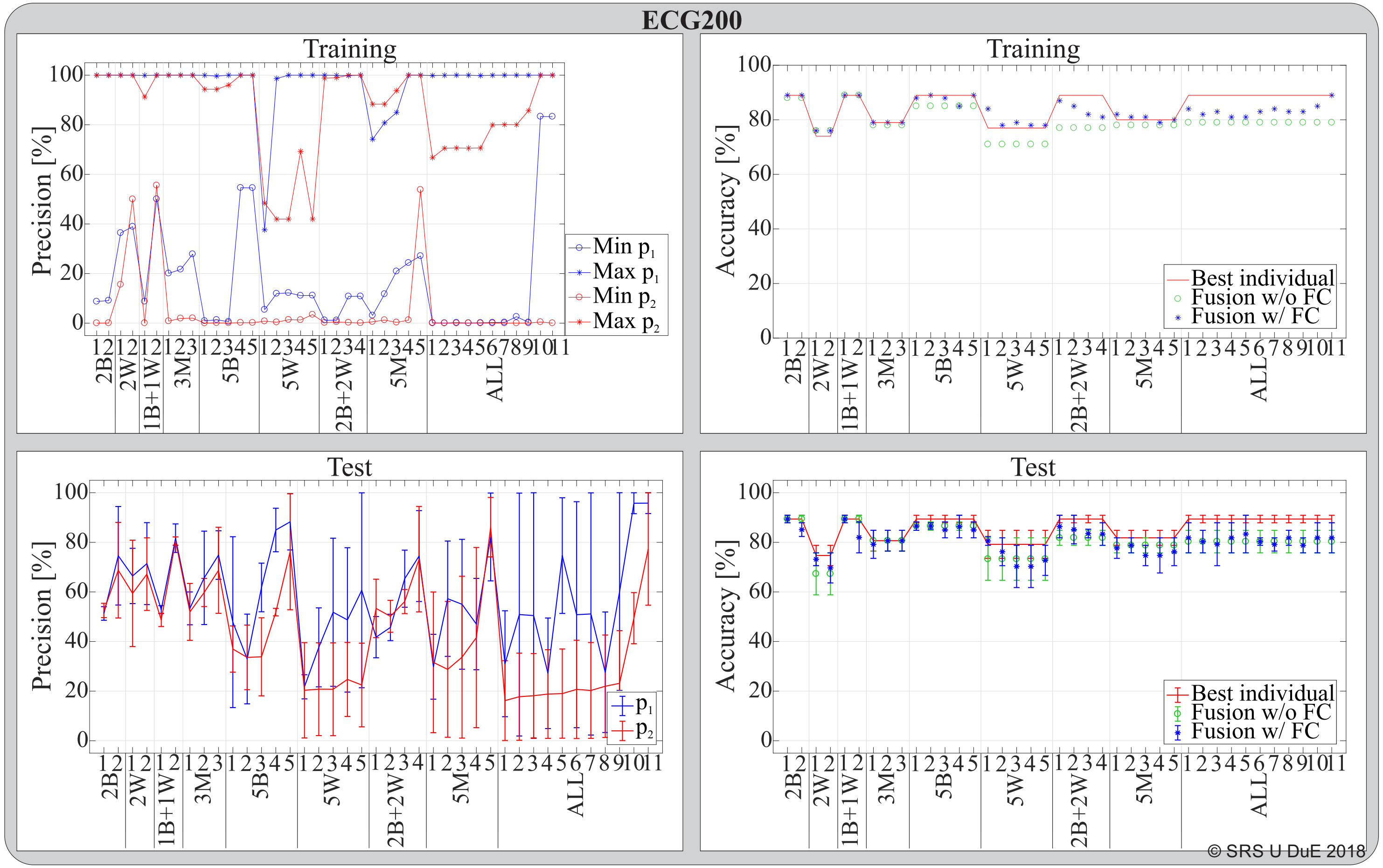

Figure 4.4: Simulation results of data set ECG200 (FC included in ensemble) [RS19].

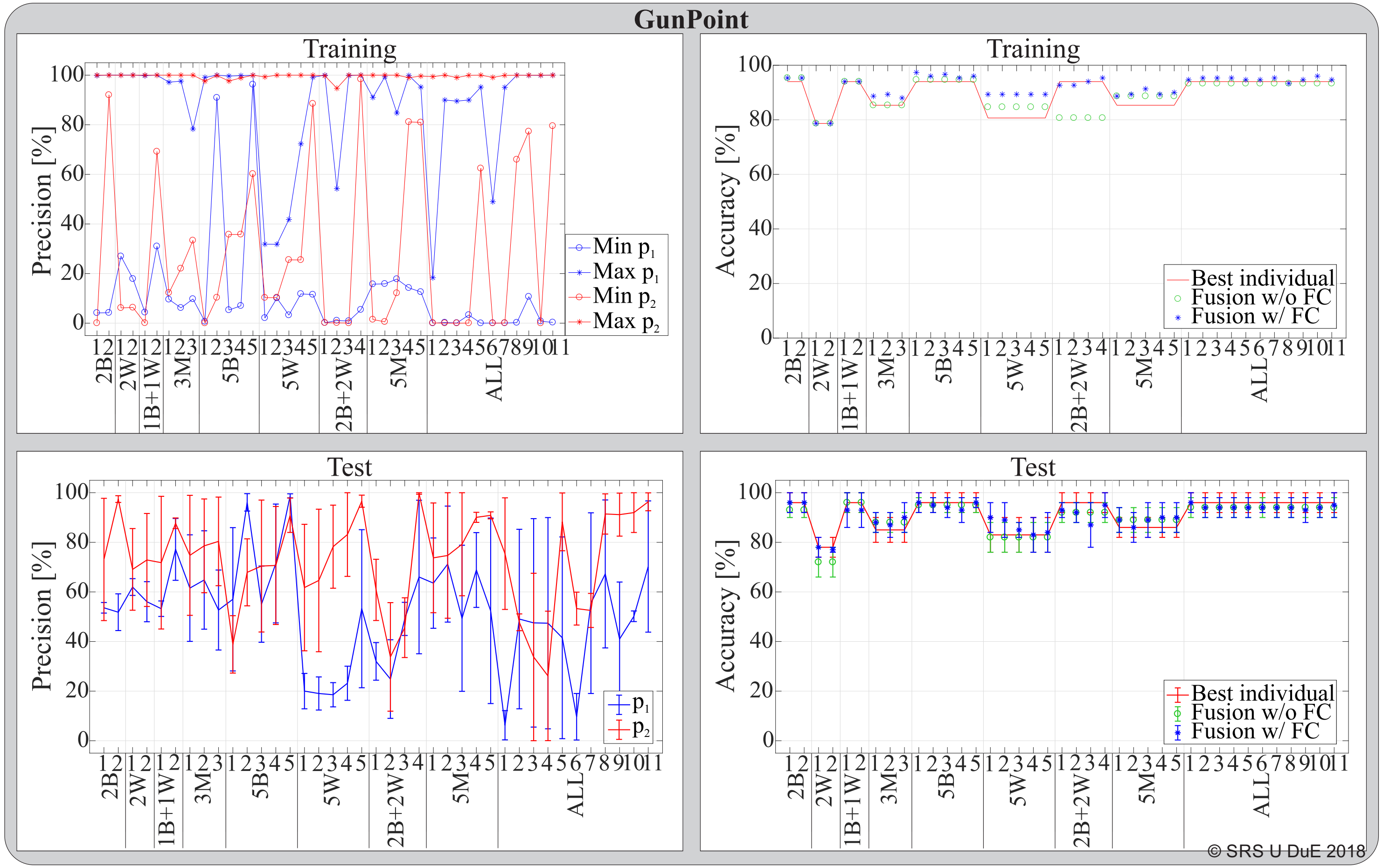

Figure 4.5: Simulation results of data set GunPoint (FC included in ensemble) [RS19].

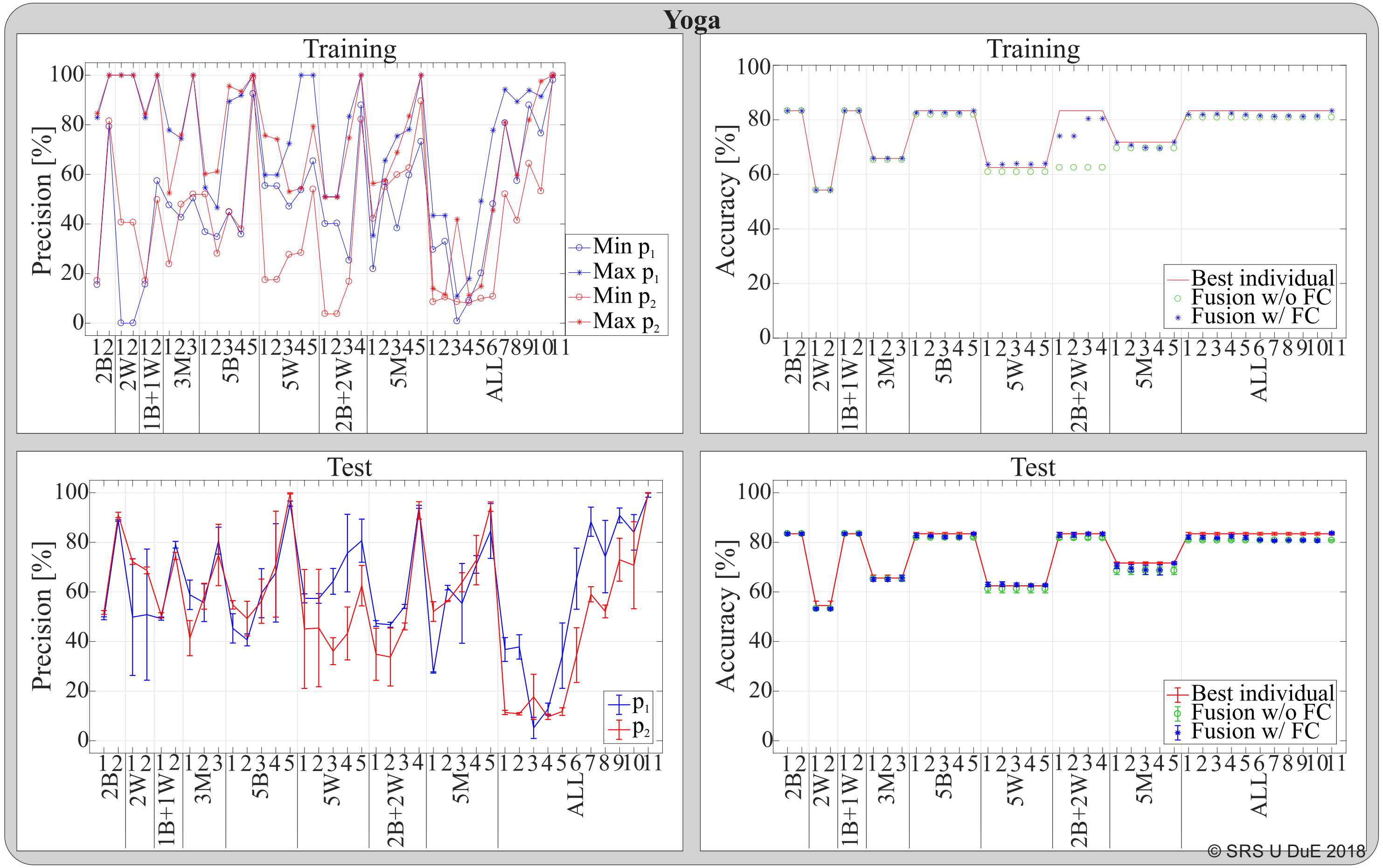

Figure 4.6: Simulation results of data set Yoga (FC included in ensemble) [RS19].

4.2.2 Additional classifier set as fictional classifier

In the previous considerations, one classifier of the different ensembles is set as FC to analyze the influences of changing precision values. In the following, the FC is an additional classifier fused with the best individual one to evaluate whether the addition of a classifier can improve the overall performance or which properties (here: precision values) the additional classifier should have to improve or not deteriorate the performance. The best individual classifier (B) is fused with each of the other classifiers separately (2-11) using BCR, where the number denotes the order according to individual accuracy (B+i, where i denotes the i^{th} best individual classifier). In the different ensembles, only two classifiers are fused, and each of them is once denoted as FC with varying precision values according to Figure 4.1 (B+i W means the worse classifier of the ensemble is set as FC, B+i B denotes the best classifier as FC).

The results of data set Coffee are shown in Figure 4.7. Using training data, the accuracy can not be improved by adding a second classifier. For all different ensembles the maximum achievable accuracy using FC and also the fusion without FC show same accuracy as the best individual classifier. From Figure 4.7 top left, the maximum and minimum precision values for each class leading to the accuracy are shown. Whenever the worse classifier is denoted as FC, the range of precision values is very high ($p_1 \approx [5\ \%, 100\ \%]$ and $p_2 \approx [0\ \%, 95\ \%]$). Independent from the precision of the worse classifier, the fused accuracy is the same for each ensemble. Considering the best classifier as FC, the range of the precision values is lower ($p_1 \approx p_2 \approx [50-70\ \%, 100\ \%]$) with small changes of the minimum values depending on the additional classifier, except of the values for the ensemble B+2 B. The two best classifiers for data set Coffee have the same individual accuracy, therefore the minimum of precision values is the same as for the ensemble B+2 W. The results from training lead to very specific selected precision values for test (see Figure 4.7 bottom left). In test results the accuracy is mostly the same for all three considerations (best individual classifier, fusion without FC, and fusion with FC). In the case of B+3 B and B+5 B, the selected precision values lead to a deterioration of performance, in case of B+11 W, the change of precision value improve the fusion performance.

Considering the training results for the data set ECG200 (see Figure 4.8 top right), again no improvement compared to best individual classifier can be achieved using fusion (with and without FC). The fusion without FC shows, that the achievable accuracy can be outperformed by changing precision values of one classifier. The improvement by changing the precision values show that the initially calculated precision values from the classifier denoted as FC are not in the given range between maximum and minimum precision values shown in Figure 4.8 top left. Similar to the data set Coffee, the range between maximum and minimum precision values is higher regarding the worse classifier as FC (B+i W) than the best classifier set as

FC (B+i B). According to this, the selected precision value for test using worse classifier as FC is in all cases around 50 % and the accuracy for test samples is the same as best individual accuracy. For the best classifier set as FC, the accuracy is in some cases the same and in some lower as the best individual accuracy, but the absolute value of deterioration is lower for fusion with FC (minimum accuracy is 87 %) than for fusion without FC (minimum accuracy is 82.5 %) (see Figure 4.8 bottom right).

Using fusion of the ensembles for training data of data set GunPoint, in Figure 4.9 top right a very low and not significant improvement of accuracy is possible. For these ensembles (B+2, B+3, B+10, and B+11), the previously explained scheme of precision values leading to this improved accuracy (high range between maximum and minimum precision value for worse classifier set as FC and low range for best classifier denoted as FC) is changed (see Figure 4.9 top left). Especially the minimum value of p_2 is higher than for the previous data sets and also for the other ensembles of this data set. This observation can not be transferred to the test results. In case of B+3 and B+11, the test results show a deterioration of fusion with FC compared to the best individual accuracy. Although in most cases, the accuracy of fusion with FC is higher than the fusion without FC.

Regarding the training and test results of data set Yoga, the accuracy of best individual classifier, fusion without FC and fusion with FC are the same for all ensembles (see Figure 4.10 top and bottom right). Adding a classifier to the best individual one does not change the performance, but the precision values leading to this maximum accuracy are in case of data set Yoga very specific. The ranges between minimum and maximum precision values are nearly the same for p_1 and p_2 and lower for best classifier as FC than for worse classifier as FC (see Figure 4.10 top and bottom left). The only change can be observed considering the minimum value of p_1 and p_2. For higher number of second classifier (means worse performance of the additional classifier), the minimum precision values decrease.

Lessons learned Adding only one classifier to the best individual does not improve the overall accuracy, but using fusion, the varying precision value lead to an improvement compared to the fusion with the initially calculated precision values. Thus, using fusion of two classifiers, the fusion using an FC lead to more robustness (with respect to the performance of the additional classifier). An additional classifier, which is worse than the other one, can have a wide range of precision values leading to the same fused accuracy. The best individual one should have specific precision values when combining with an additional worse classifier.

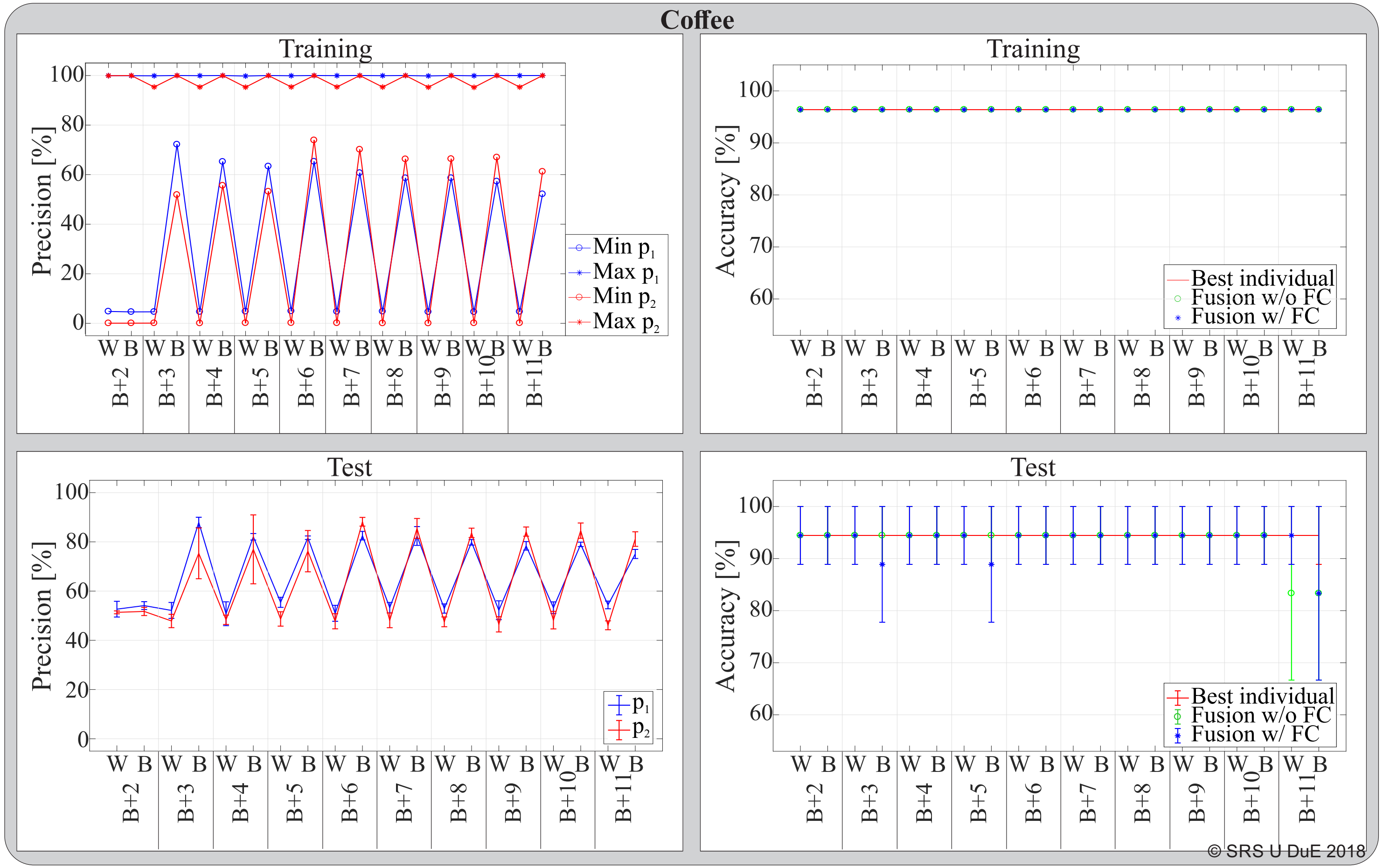

Figure 4.7: Simulation results of data set Coffee (adding an FC).

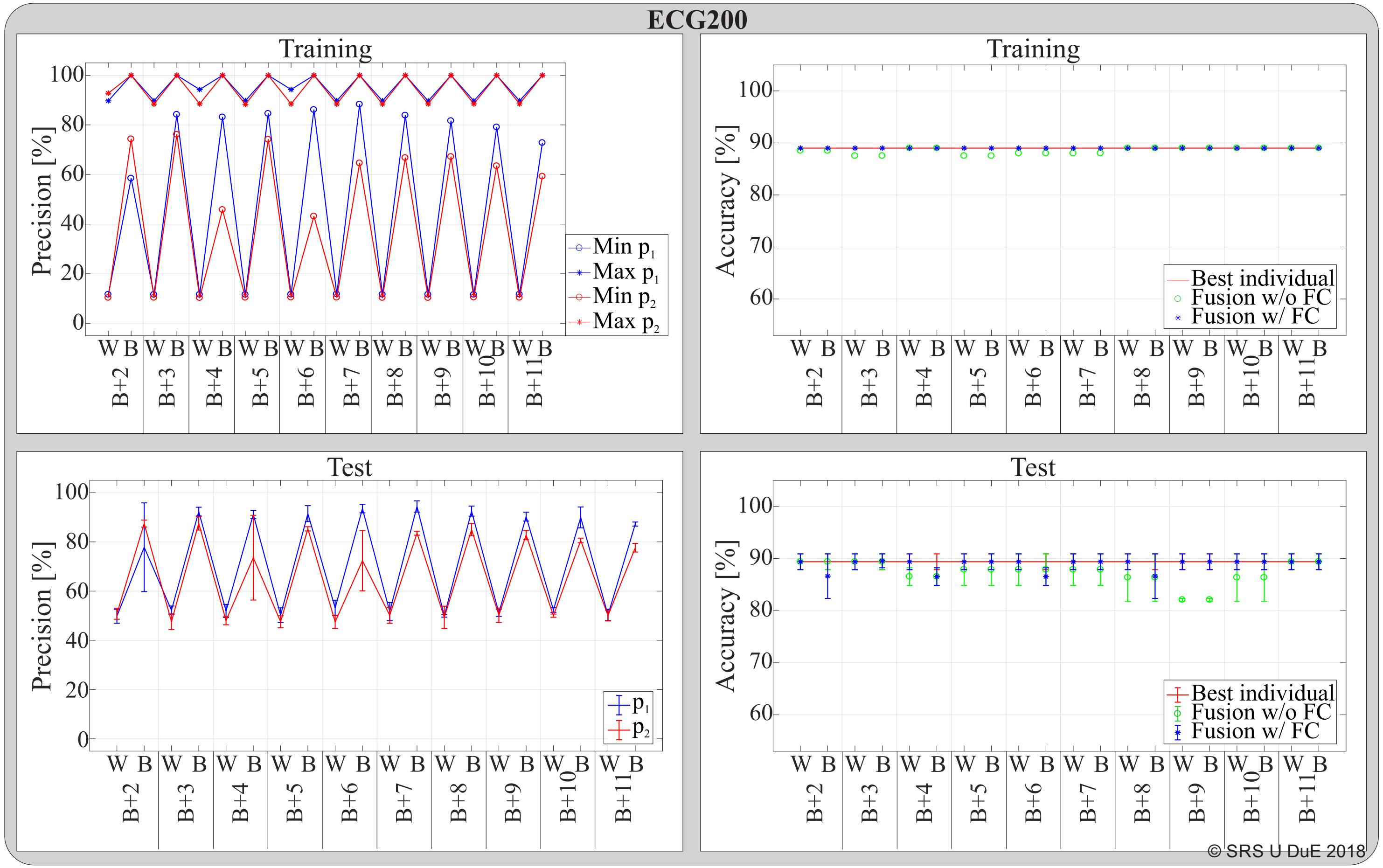

Figure 4.8: Simulation results of data set ECG200 (adding an FC).

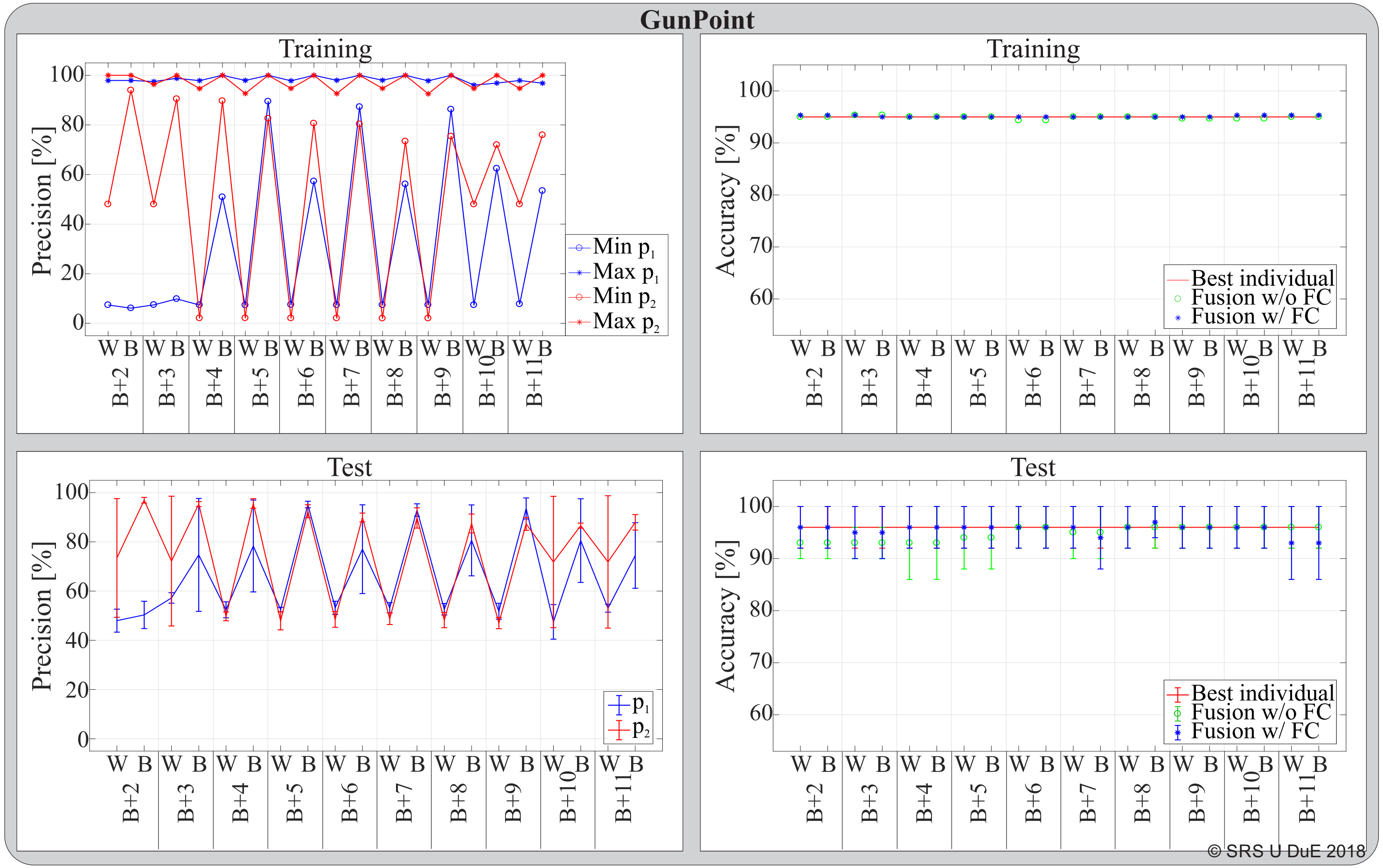

Figure 4.9: Simulation results of data set GunPoint (adding an FC).

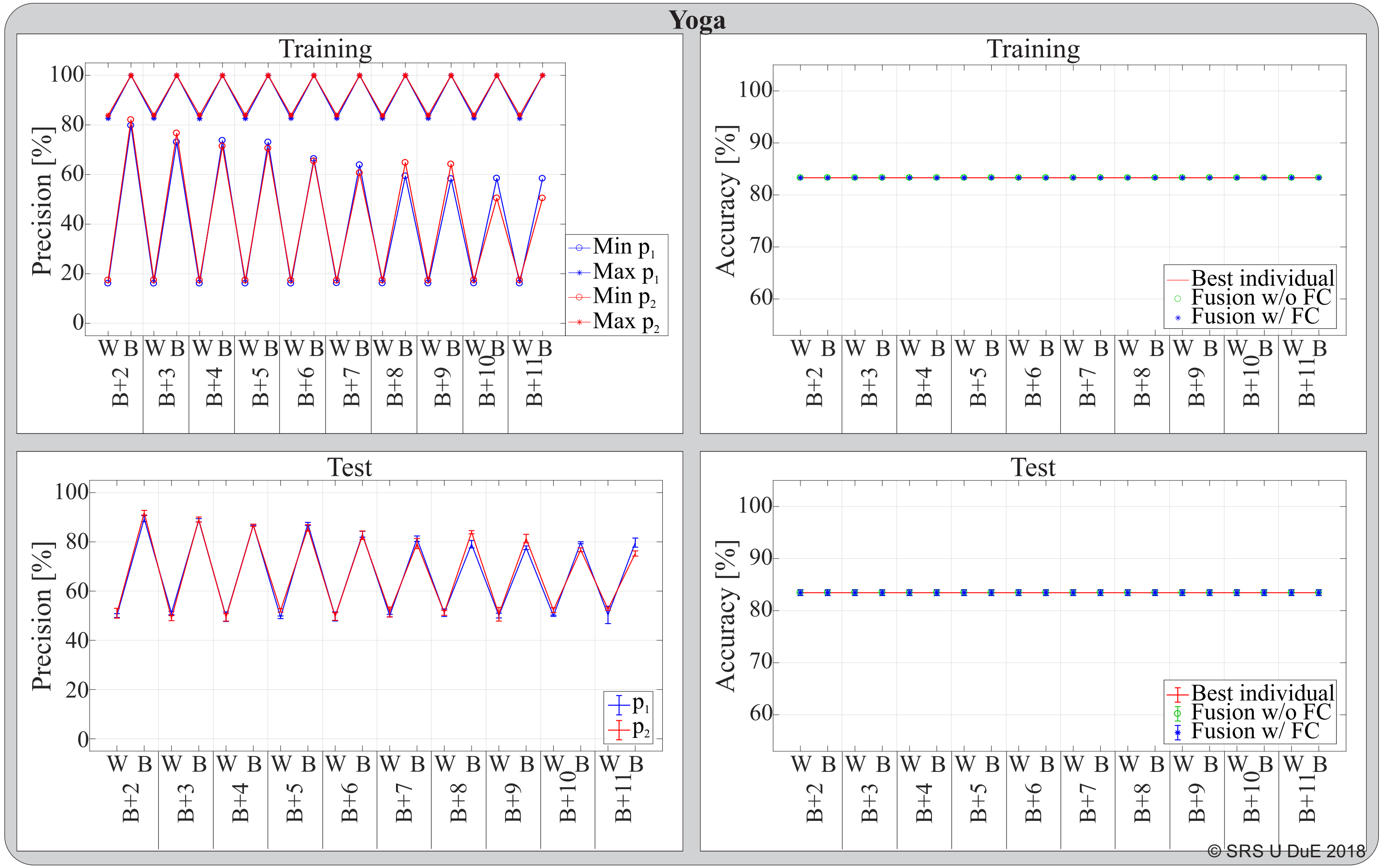

Figure 4.10: Simulation results of data set Yoga (adding an FC).

4.3 Application to experimental data

The idea of fictional classifier is applied to real industrial data from hot rolling mills (HRM). First results using FC regarding the distinction of four classes are shown. The background of the data, like classes, used classifiers, and individual classification results are shown in Chapter 3.3.1.

In the following 3-fold cross-validation is used to generalize the results. In Tab. 4.2 the used static ensemble strategies are shown. Considering a four class problem, the probability matrix contains of 16 entries. Due to the requirement, that the sum of each column in the probability matrix has to be equal to 1, the number of parameters to optimize is 12 (P_{ij}^{FC} with $i = 1, \ldots, n_C - 1$ and $j = 1, \ldots, n_C$). The probability matrix results in

$$P^{FC} = \begin{bmatrix} P_{11}^{FC} & \cdots & P_{14}^{FC} \\ P_{21}^{FC} & \cdots & P_{24}^{FC} \\ P_{31}^{FC} & \cdots & P_{34}^{FC} \\ 1 - \sum_{i=1}^{3} P_{i1}^{FC} & \cdots & 1 - \sum_{i=1}^{3} P_{i4}^{FC} \end{bmatrix}. \tag{4.2}$$

This means, that not only the precision values ($i = j$) are variable, but also the probability the classifier assigns class i, but the real class is j ($i \neq j$).

In Figure 4.11 the accuracies of best individual classifier, fusion without and with FC of training and test process are shown. Compared to the results of benchmark data with two classes, the training results show different behavior. Here an improvement using fusion is always possible. Using the FC, the accuracy is in some cases higher in some cases less than without FC depending on the applied ensemble. There are also two cases (2W 2 and 1B+1W), in which the results using FC are significantly worse compared to the best individual accuracy and the fusion without FC, especially when the worst classifier in the ensemble is set as FC. In test the accuracy using fusion with FC is not improved compared to fusion without FC in most cases. In general the results using fusion (with or without FC) can not outperform the best

Abbreviation	Ensemble selection
2B/3B/4B/5B	Two/three/four/five best classifiers
2W/3W/4W/5W	Two/three/four/five worst classifiers
1B+1W/2B+2W	One/two best and one/two worst classifiers
2M/4M	Two/four medium-good classifiers
ALL	All six classifiers

Table 4.2: Static classifier ensembles selected by accuracy of individual classification results applied on experimental data [RS19].

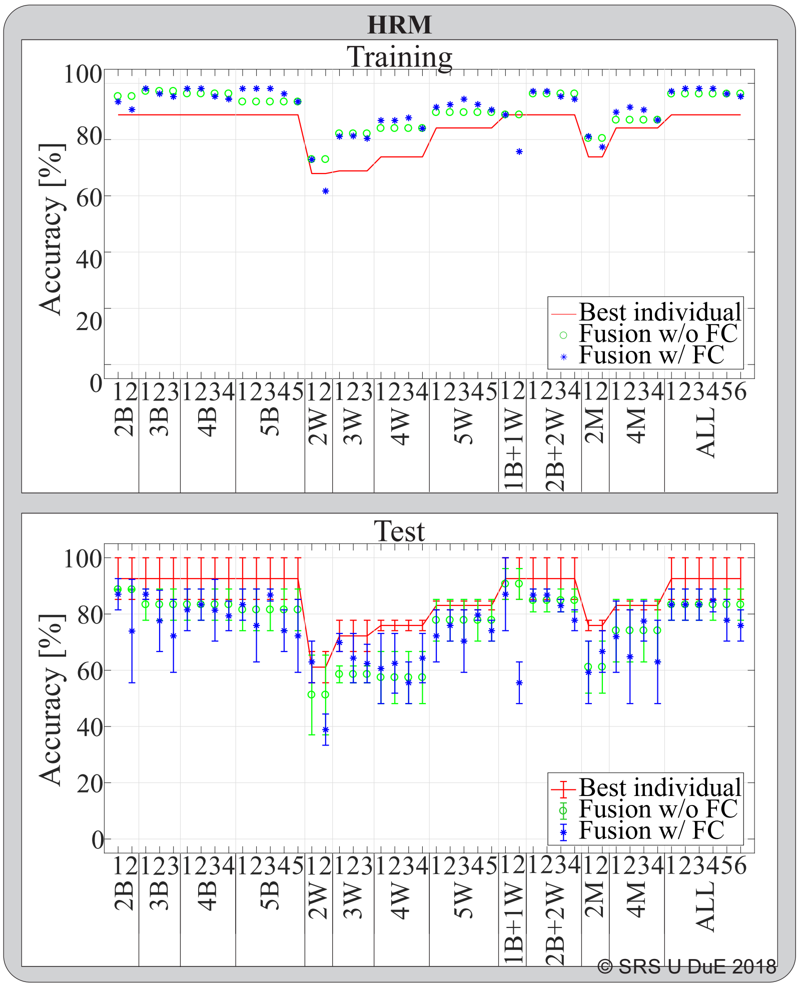

Figure 4.11: Simulation results of data set HRM [RS19].

individual classifier. Obviously the values calculated/optimized during training can not be applied for test data.

Lessons learned Compared to the results from benchmark data, the experimental results lead to similar conclusions with some restrictions. In training the accuracy can be improved using the optimized parameters for specific ensembles and classifiers set as FC. The improvement potential is different for all considered cases. Considering a four class problem, the optimized parameters during training can not be directly transferred to unknown samples. Using the fictional classifier the improvement potentials also for a four class problem can be shown. In this case the optimized 12 parameters are too specific to select them for unknown samples.

4.4 Conclusions from the analysis of classifier properties

The questions regarding the properties a classifier of an ensemble should have as well as whether the use of a fictional classifier can lead to better fusion performance are considered in this chapter. Resuming all example combinations it can be concluded that in general a higher accuracy can be obtained using the FC compared to the accuracy of fusion without FC. The best individual classifier can not be outperformed, although for specific ensembles (those selecting the worst classifiers) the analysis using FC show an improvement potential varying the precision values.

The precision values an additional classifier should have can also be stated from the analysis using the FC. An additional classifier, which is worse than the other one, can have a wide range of precision values leading to the same fused accuracy. The best individual one should have specific precision values when combining with an additional worse classifier.

How can this result be used?

- There is no best classifier performing best for all data sets. This is a known result denoted as 'No Free Lunch Theorem' [WM97].
- Knowing the specific conditions for classification it makes sense to search for the best individual classifier to perform best in specific applications.
- Not knowing specific conditions, fusion is useful while providing more robust (with respect to new and untrained situations) results.
- Using fusion, it is useful to evaluate the improvement potentials by varying the precision values and determine the best ensemble and classifier as FC.
- If no improvement is possible, common fusion methods (like the Bayesian Combination Rule) can be used.
- If an improvement is possible, the precision values as well as the ensemble selection and classifier set as FC from training can be used for fusion of new samples.
- If only classifiers with worse performance are available, a fusion with FC should always be considered.

The evaluation using the introduced fictional classifier can help to analyze the improvement potentials and assists by deciding if and under which conditions a fusion of results is useful.

5 Variation of performance measures used for fusion

As explained in the previous chapter, the precision value is an influencing factor to the overall accuracy of fused results. In some applications, the precision value is not computable because of not available training or validation data. Probability estimations are available as classifier output denoting the probability of a sample belonging to the assigned class. This value can directly replace the precision value, because both are probabilities belonging to one class. Furthermore, conventional NDT approaches use the Probability of Detection (POD) as a reliability measure instead of precision. The POD quantifies a sensor-filtering approach with respect to mostly static measurements and are dependent on the flaw size a, so the replacement of the precision values with the POD values is more complex. This chapter details a new concept replacing the precision values with POD values and shows experimental examples for condition monitoring using the new concept or replacing the precision value with the probability estimations respectively.

In Chapter 5.1 the new approach used to improve the POD characterization of each sensor-/vibration-based statement by decision fusion using several sensors is introduced. To fuse the detection results of the sensors related to their POD and $a_{90/95}$ value, the BCR is used. The application to two experimental examples is explained in Chapter 5.2. In the first experimental example, the POD values for fault diagnosis in elastic structures are calculated and used for fusion. The second example shows the replacement of the precision values using the probability estimations. Conclusions are given in Chapter 5.3.

The content, figures, and tables in this chapter are based on publication of [ARS18c], [ARS18a], [ARS18b], and [Rot^{+}17] and prepared for publication of [ARS19].

5.1 Concept of POD-based fusion

In the case of fault detection, normally the precision value is used as a performance measure, which is considered in the fusion process. Here the measurable POD values for specific masses can replace the precision value, because both define a performance measure about the reliability of an assignment. Therefore, the POD of each sensor and feature for specific faults (here: masses denoted as $a_{90/95}$ values for the considered sensor-feature combination) can be used to calculate the belief values according to the Bayesian combination rule. The procedure is shown in Figure 5.1. First the POD curves for all n sensor-feature combinations are calculated. From the POD curve, the $a_{90/95,i}$ value with $i = 1, ..., n$ for all n combinations can be determined. Corresponding to each $a_{90/95,i}$ value, one POD value ($\mathrm{POD}_j(a_{90/95,i})$ with $j = 1, ..., n$) can be assigned using the calculated POD curve (in total n times

n POD values). For further considerations, the $a_{90/95,i}$ values are treated like classes, where as for unknown situations, each sensor-feature combination j can just detect a fault or not. Based on the detection, the precision for one $a_{90/95,i}$ value P_j used for fusion is set as $P_j(a_{90/95,i}) = \mathrm{POD}_j(a_{90/95,i})$ in case of detection and $P_j(a_{90/95,i}) = 1 - \mathrm{POD}_j(a_{90/95,i})$ in case of no detection. Using the standard Bayesian Combination Rule, the values are combined to one belief value (here: one value for each $a_{90/95,i}$ value). Using the belief values, for each detection combination, one belief-curve can be calculated.

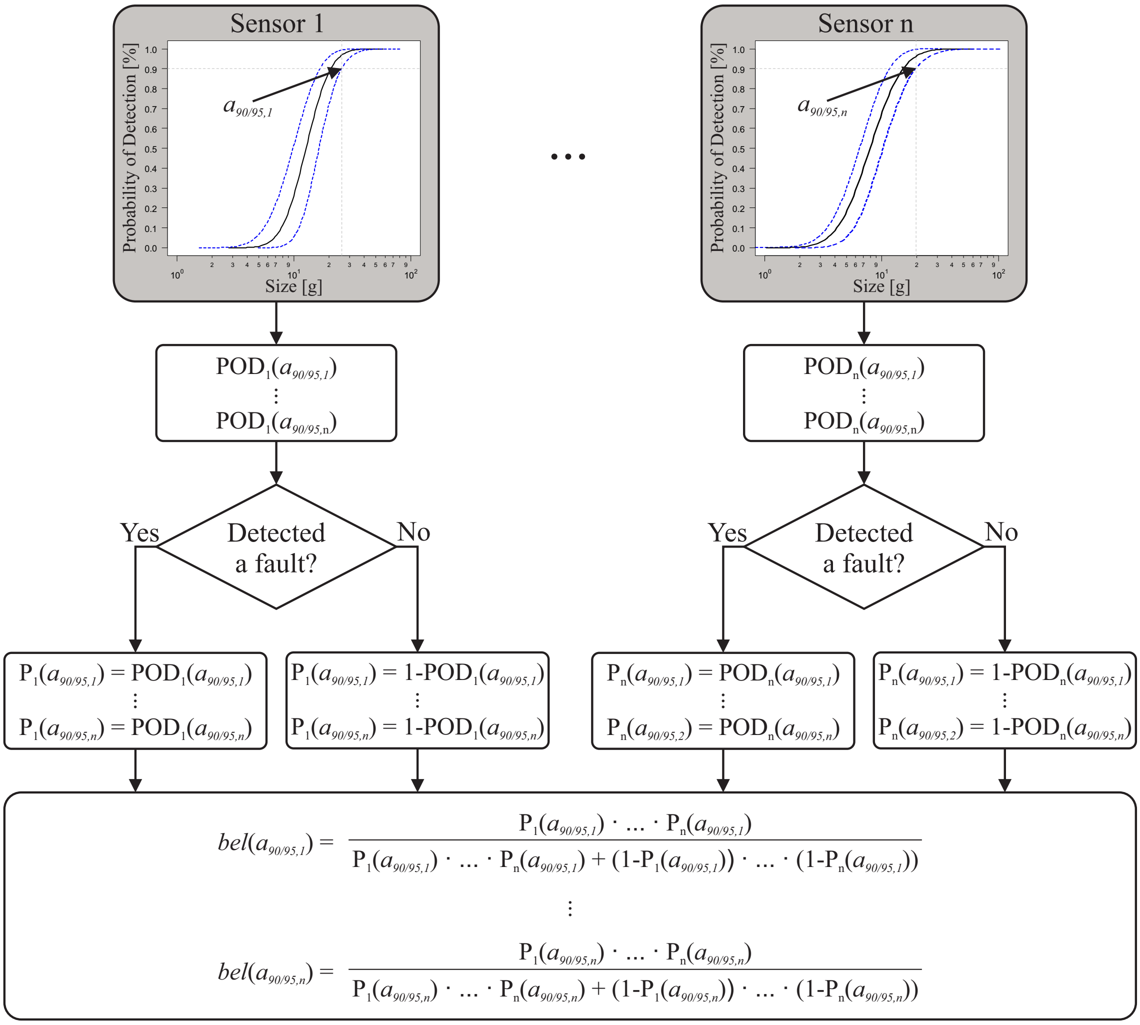
Sensor 1
Probability of Detection [%]
Size [g]
a90/95,1
...
Sensor n
a90/95,n
POD1(a90/95,1) ⋮ POD1(a90/95,n)
PODn(a90/95,1) ⋮ PODn(a90/95,n)
Yes
Detected a fault?
No
P1(a90/95,1) = POD1(a90/95,1) ⋮ P1(a90/95,n) = POD1(a90/95,n)
P1(a90/95,1) = 1-POD1(a90/95,1) ⋮ P1(a90/95,n) = 1-POD1(a90/95,n)
Pn(a90/95,1) = PODn(a90/95,1) ⋮ Pn(a90/95,2) = PODn(a90/95,n)
Pn(a90/95,1) = 1-PODn(a90/95,1) ⋮ Pn(a90/95,2) = 1-PODn(a90/95,n)
bel(a90/95,1) = P1(a90/95,1) · ... · Pn(a90/95,1) / (P1(a90/95,1) · ... · Pn(a90/95,1) + (1-P1(a90/95,1)) · ... · (1-Pn(a90/95,1)))
⋮
bel(a90/95,n) = P1(a90/95,n) · ... · Pn(a90/95,n) / (P1(a90/95,n) · ... · Pn(a90/95,n) + (1-P1(a90/95,n)) · ... · (1-Pn(a90/95,n)))

Figure 5.1: Concept of POD-based fusion.

5.2 Application to experimental data

The concept for POD-based fusion is applied to the fault diagnosis of an elastic beam. Vibration-based SHM operates on the principle that structural defects result in changes in dynamical properties. Fault identification is mainly based on displacement, velocity, or acceleration measurements at a single point [Fri05]. Therefore, different sensors and extracted features are considered with different performance expressed by $a_{90/95}$ values. The replacement of probability estimations for precision values is shown using experiments of damage detection in composites.

5.2.1 Fault diagnosis in elastic structures

The experimental system to be considered for illustration is an elastic beam. Acceleration, displacement, and strain measurements are taken. As features, band power and eigenfrequency analysis are carried out on the first two modes of the mechanical system.

The experiment is carried out on an elastic mechanical beam using a test rig. An elastic steel beam of dimensions 545 x 30 x 5 mm is clamped on one side. The beam length is divided into five equal parts (Figure 5.2) defining sensors position. Piezoelectric accelerometers (ACC) are attached at three positions (P1, P2, and P3) on the beam. Two strain gauges (SG) are bonded onto the beam at positions P1 and P3. Two displacement measurements are taken at the two positions (P2, P4) using non-contact laser sensors (Laser). The beam can be excited manually or by modal hammer. Changes within the elastic mechanical structures are assumed as changes due to varying mass, so here additive masses are applied to modify the existing initial system to simulate a fault (due to mass change). Two cases of point mass placement are examined. Case I involves placing the point mass at midpoint of position 2 and 3. Case II involves the placement of point mass at the midpoint of position 3 and 4. These masses are added to the specified locations. For every incrementally placed mass the beam is excited and the corresponding data are

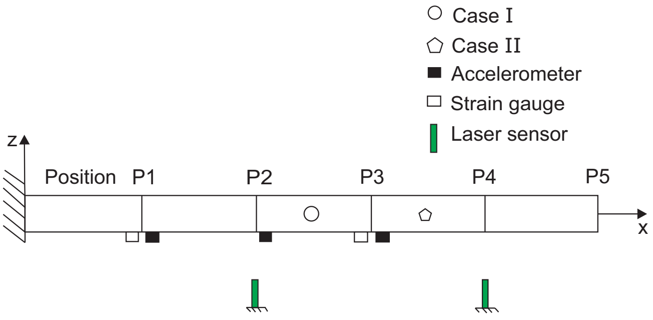

Figure 5.2: Sensor positions relative to beam length [ARS18a].

recorded. The analysis is carried out for the first and second mode for each situation of mass placement (cases I and II). To explore the non-uniqueness of a feature for POD analysis, as additional feature, band power is also extracted. In Table 5.1 the results of the $a_{90/95}$ values for individual sensors and features are given. The best results for one feature are marked in bold (lowest $a_{90/95}$ value) and the worst results are underlined (highest $a_{90/95}$ value). From the results, it becomes evident that the $a_{90/95}$ POD quantification is different depending on sensor type and feature considered. The numbers indicate the maximum mass that can be missed with a 90 % POD at a 95 % confidence level. Lowest masses represent best results, so the sensor detects least mass change. Worst POD sensor characterization, represent worst results so the sensor requires large fault (here: mass) values to be detected with $a_{90/95}$ reliability.

In the case of non reliable statements (the use of one sensor and feature is low to ensure confident health status statements) decision fusion of the results from different sensors or features may be an option for improving reliability.

In Table 5.2 the POD of the seven sensors for the corresponding $a_{90/95}$ values calculated from POD curve shown in Figure 5.3 are given. The 90 % values are marked in bold to show the correspondence between the sensor and the related $a_{90/95}$ value. The POD values for masses lower than the individual $a_{90/95}$ value are smaller, than those for higher masses.

To explain the combination using POD values, the fusion of the sensors ACC 1 and ACC 2 and the feature band power is considered as example. To calculate the belief values, it has to be known, which sensor detected the fault and which failed. Assuming ACC 1 detected a fault, ACC 2 did not, the precision values are denoted

Table 5.1: Resulting $a_{90/95}$-values [g] of the different measurements [ARS19].

	Point mass between P2 and P3 (Case I)			Point mass between P3 and P4 (Case II)		
Sensor	Mode 1	Mode 2	Band power	Mode 1	Mode 2	Band power
ACC 1	74.04	**48.15**	45.28	52.21	9.915	**84.56**
ACC 2	74.04	55.78	34.63	52.15	9.293	22.78
ACC 3	74.04	72.59	**20.20**	52.15	13.36	**17.29**
SG 1	85.19	72.38	29.34	52.15	11.07	28.69
SG 2	126.70	62.23	34.37	54.08	9.394	27.93
Laser 1	**67.30**	61.03	27.64	**52.10**	43.49	25.54
Laser 2	74.04	-	23.44	52.15	-	25.85

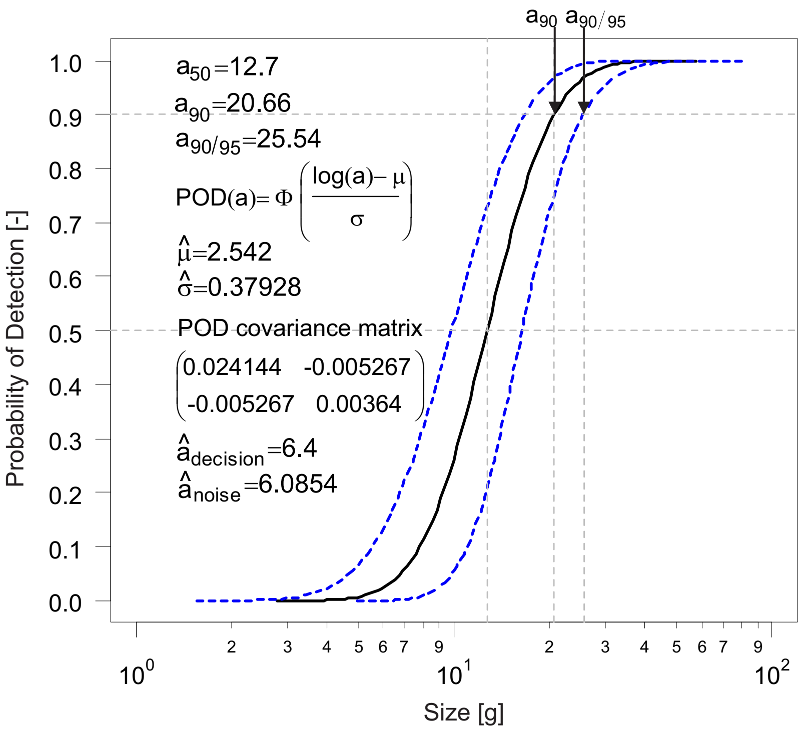

Figure 5.3: POD related to Laser 1 and band power as feature [ARS19].

as $P_1 = \text{POD}_{\text{ACC1}}(45.28\text{ g})$ and $P_2 = 1 - \text{POD}_{\text{ACC2}}(45.28\text{ g})$. The belief value for the mass of 45.28 g ($= a_{90/95,ACC1}$) is calculated by

$$bel(45.28\text{ g}) = \frac{P_1 \cdot P_2}{P_1 \cdot P_2 + (1 - P_1) \cdot (1 - P_2)} \ . \tag{5.1}$$

In the same way, the belief values for the other masses can be calculated. Extending the fusion to all sensors, the number of detection combinations ascends.

The resulting belief values for different detection combinations, as example if five of the seven sensors detected a fault (e.g. 0 0 1 1 1 1 1 means the first two sensors (ACC 1 and ACC 2) did not detect a fault, all others did), are shown in Figure 5.4. Depending on which five of the seven sensors detect the fault, the belief values vary. Although all belief values increase for increasing masses. This means that the probability, that a fault with a higher mass is present is higher in case of five sensors detecting a fault. In Figure 5.5 a selection of different detection combinations is presented. For selection of the best and worst sensor-/feature-based statement, results according to Table 5.1 are considered. For example 4B means the four best sensors have detected a fault, the others did not. In case of 6B, 5B, and 4B, all belief values are close to 100 %, whilst in case of 1W, 2W, and 3W are close to 0 % (see Figure 5.5). For the other cases a symmetry can be seen, e.g. the curve of 5W corresponds to 100 % minus the curve of 2B.

Table 5.2: POD results [%] for different $a_{90/95}$ values (only for band power as example) [ARS19].

Sensor	20.20 g	23.44 g	27.64 g	29.34 g	34.37 g	34.63 g	45.28 g
ACC 1	48.6	61.8	71.6	74.0	83.9	84.8	**90.0**
ACC 2	63.5	71.2	83.2	84.8	89.7	**90.0**	96.1
ACC 3	**90.0**	95.0	97.1	97.6	98.8	98.9	99.5
SG 1	74.5	83.9	85.8	**90.0**	94.4	94.7	95.8
SG 2	64.2	75.5	84.3	85.7	**90.0**	91.1	96.0
Laser 1	77.7	87.1	**90.0**	91.6	94.5	95.4	98.1
Laser 2	85.5	**90.0**	93.7	95.3	97.2	98.1	99.2

Considering the case 2B, which means the two best sensor-/feature-based statement (with the lowest $a_{90/95}$ value) are detecting a fault, all other 5 sensors are not, it is more probable, that there is a small fault than a bigger one, because if the better

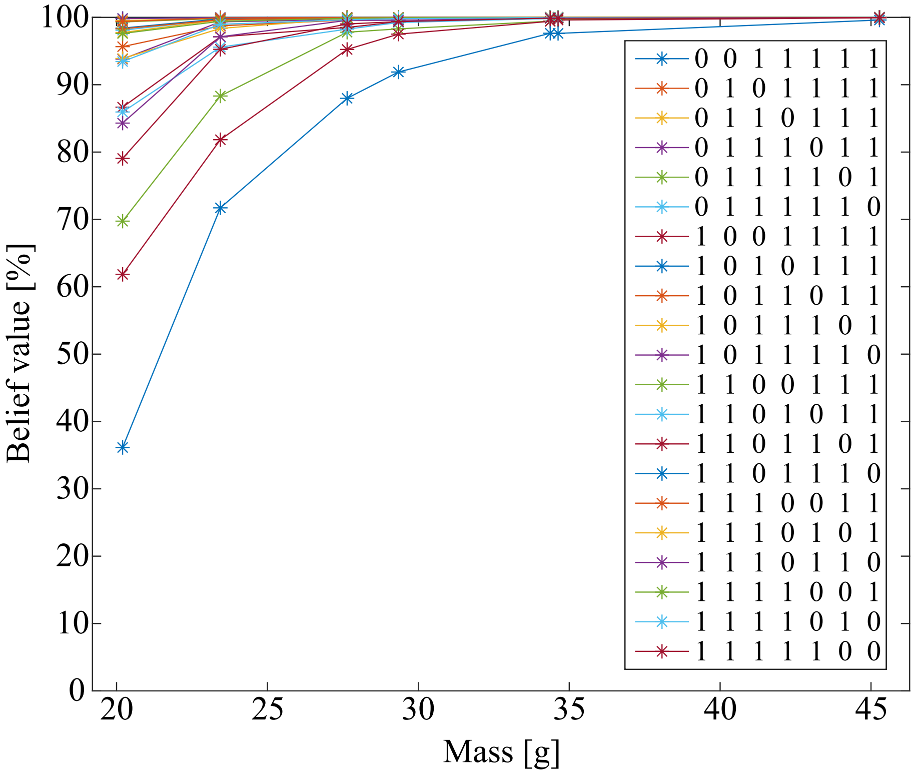

Figure 5.4: Belief values for different detection combinations, when 5 of 7 sensors detect a fault [ARS19].

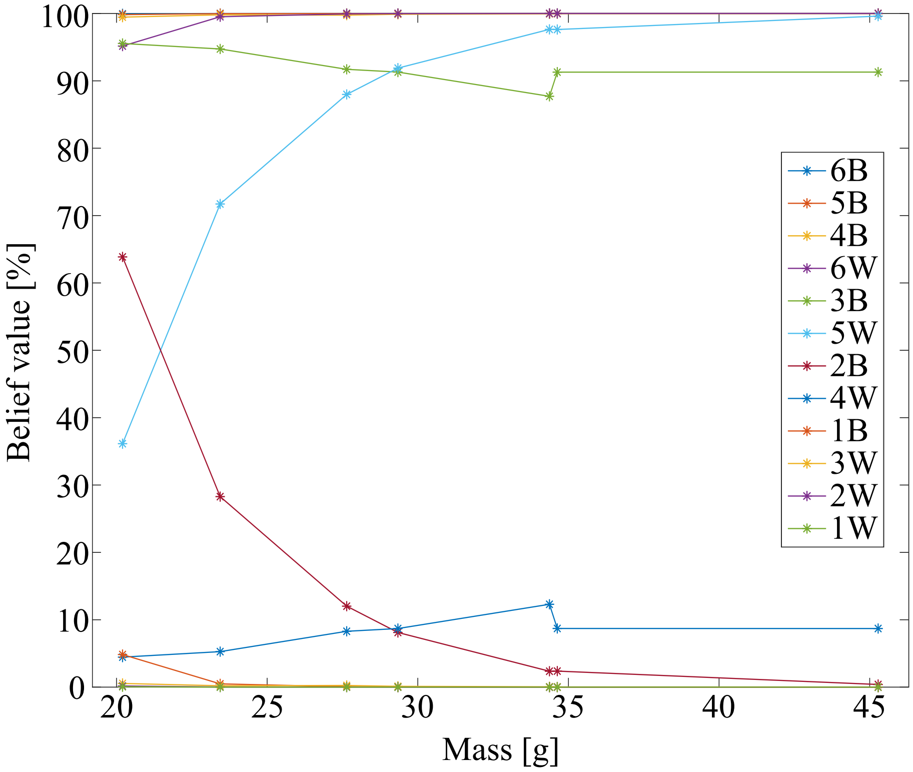

Figure 5.5: Belief values for different detection combinations, selected by the best (B) or worst (W) sensors detecting the fault [ARS19].

sensor-/feature-based statement have detected a fault, it could be possible, that the fault is too small to be detected by the other sensors (with higher $a_{90/95}$ value). However, if the mass would be larger, the other sensors should also have detected the fault. If there is a higher number of sensors detecting a fault (like in case 5W), the belief values increase for increasing masses, because a larger fault is easier to detect.

Lessons learned The combined POD curves resulting in one fused curve denote the belief value for specific fault sizes (here masses). Depending on which detection systems detected a fault, the probability, that a fault with a specific fault size exists, increases or decreases with increasing fault size. In case of decreasing belief values, it is more probable, that there is a small fault than a bigger one, because the better sensor-/feature-based statement have detected a fault. The fault can be too small to be detected by the other sensors with higher $a_{90/95}$ value. The belief values increase for increasing masses, if there is a higher number of sensors detecting a fault (like in case 5W). A larger fault is easier to detect, so that more even worse detection systems are detecting the fault.

5.2.2 Damage detection in composites

The injection of damages in composites is more difficult than for other materials like the addition of masses to the elastic beam. Using the data presented in [WBS18], the measurement series of each specimen comprises a total of 25 data sets. The appeared damages in these experiments are not known. Measurements were performed using 5 different excitation frequencies in the range of [2 Hz, 6 Hz] and 5 different amplitudes in the range [6 mm, 18 mm]. Using Acoustic Emission measurements the data is classified by Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Artificial Neural Network (ANN).

The classification performance is assessed using probability estimation indicating the likelihood that a particular observation is a member of the assigned class (delamination, matrix crack, debonding, and fiber breakage).

From the classification results, mean values of probability estimation were computed to investigate dependencies between classification reliability and loading conditions. In Figure 5.6, the loading pattern is plotted on the x- and y-axes, respectively. Corresponding probability estimates are indicated by the color scale. For each classifier (KNN, SVM, and ANN) as well as for each class (delamination, matrix crack, debonding, and fiber breakage) the probability estimations are shown depending on the excitation. The experimental results show that the classification performance varies strongly with the excitation.

The precision values used for fusion can not be calculated for this data set, because of the missing labels of the data samples. To fuse these classification results, the probability estimations are used instead of the precision values.

Considering the fusion results of WV, the probability estimation has similar values and similar distribution as for the individual classification. Using BCR, the values change to over 80 % or 0 %. This means, that if there is a decision for one class, the belief value of this class is over 80 % and therefore a crisper decision for one of the classes can be obtained. The variation of the probability estimations with the loading conditions is less significant.

Furthermore, the structure of the probability estimation distribution of the fused results shows similarities to all individual classifiers such as the three areas of class debonding, where the probability estimation is 0 % (blue). The area corresponding to the amplitude of 6 mm is also existent for classifiers KNN, SVM, and in parts of ANN. For amplitude of 12 mm and frequency of 6 Hz, there is a value of 0 % in case of classifier KNN and SVM. The probability estimation for amplitude of 18 mm and a frequency of 4 Hz is also 0 % in the results of classifier SVM and ANN.

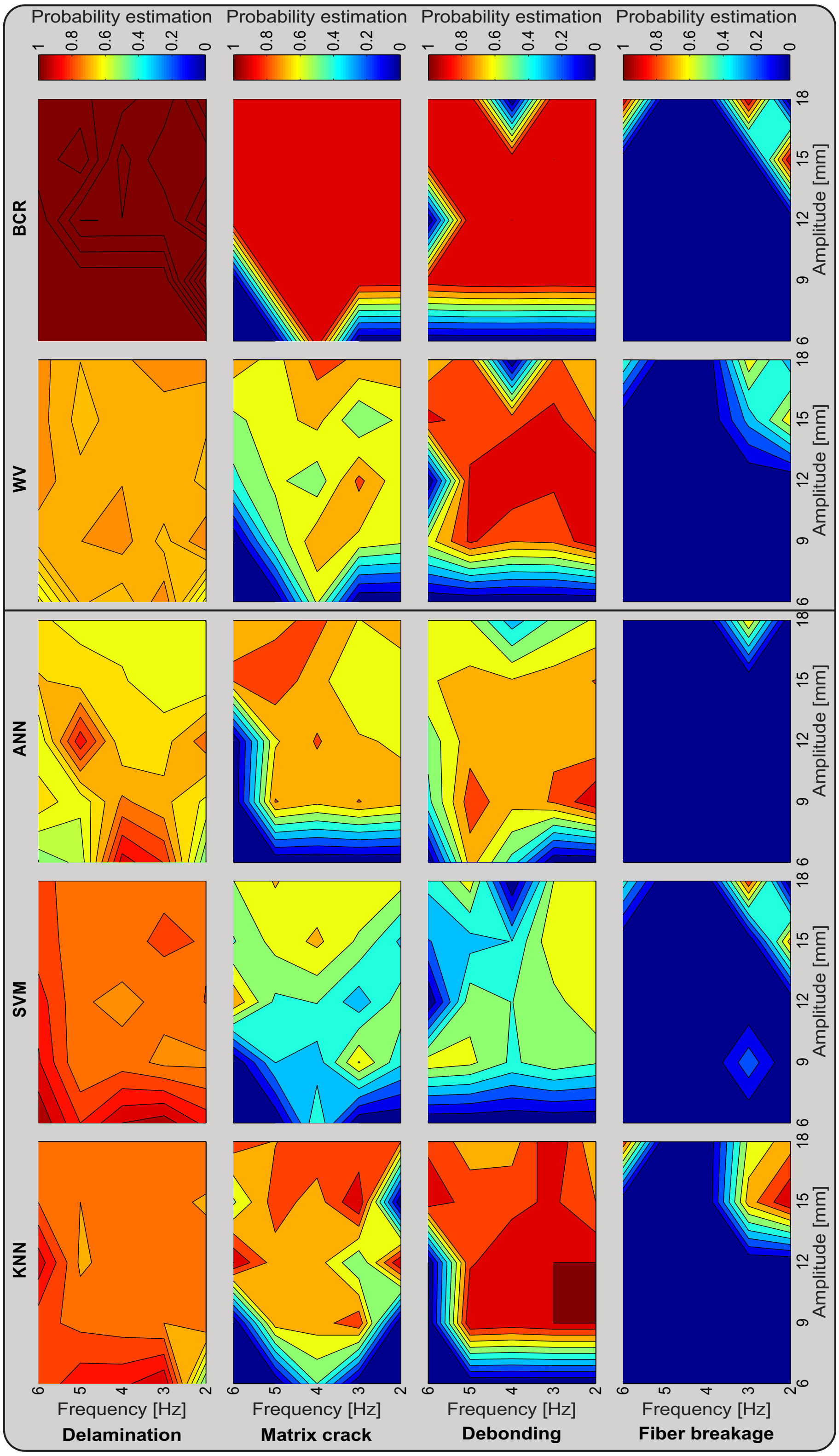

Figure 5.6: Probability estimations of classification and fusion results for different load conditions [Rot+17].

Lessons learned Fusion using the probability estimations instead of precision values is possible. Using the BCR method, the belief value leading to a decision for one of the classes is either higher than 80 % or 0 %, so that a crisper decision can be made. The dependancy of the probability estimations to the loading conditions (frequency and amplitude) is less than for individual classification.

5.3 Conclusions from the variation of performance measures

Using different performance measures (here POD and probability estimations) the fusion methods have to be adapted. The experimental examples show, that reasonable fusion results are computed using the fusion with the BCR by replacing the precision value. Several individual performances and assignments can be fused. The results of fusion using probability estimations show that fusion leads to crisper decisions whether for one or the other class.

6 Analysis of data and fusion method characteristics influencing fusion performance

In this chapter fusion options are discussed to overcome the problem of not improving the performance using fusion and to define influencing factors on overall fusion accuracy. As a result requirements for good or guaranteed or possibly increased fusion performance and also suggestions denoting those options not leading to any kind of improvement are given. For illustrating the effects, a practical example based on three characteristics of fusion methods (type of classifier output, use of these outputs and necessity of training) and four data properties (number of classes, number of samples, entropy of classes and entropy of attributes) are considered and analyzed with 15 different benchmark data sets, which are classified with eight classification methods. The classification results are fused using seven fusion methods. From the discussion of the results it can be concluded, which fusion method performs best/worst for all data sets as well as which fusion method characteristic or data property has more or less positive/negative influence on the fusion performance in comparison to the best base classifier. Using this information, suitable fusion methods can be selected or data sets can be adapted to improve the reliability of decisions made in complex or safety critical systems.

In Chapter 6.1 the selected data sets, classifiers, and fusion methods as well as the numerical analysis procedure are explained. In Chapter 6.2 the results of the numerical calculations based on benchmark data with respect to the overall performance, the fusion method characteristics, and the data properties are shown and discussed. Conclusions are given in Chapter 6.3.

The content, figures, and tables in this chapter are prepared for publication of [RKS19].

6.1 Concept of numerical analysis

To evaluate the influences of the above mentioned characteristics on the fusion performance, various experiments are conducted. Experiments instead of analytical calculations are used, because the final fusion result can not be calculated without specific assignments from the classifiers. These classifier assignments also depend on considered data sets [AS06], so that several assumptions and dependencies have to be considered. The experimental evaluation offers the advantage, that these assumptions and dependencies are fixed for the given benchmark data sets. For example-related generalization training and test data sets are divided by nested cross-validation. By means of the experiments, conclusions and suggestions to the suitable fusion methods can be given.

The specification of considered fusion methods and base classifiers, as well as the definition of the applied data sets and their fundamental properties is introduced first. The conducted experimental procedure as well as the intended purpose of several experimental design decisions will be detailed.

6.1.1 Data sets

A large amount of potentially useful measures are known to describe data characteristics. To cover a preferably wide range of characteristics, the conducted experiments are based upon 15 different problems taken from the UEA & UCR Time Series Classification Repository [Bag+17]. Selected data sets and corresponding characteristics are listed in Table 6.1. All selected data sets are characterized by nominal class labels and continuous valued numerical attributes without any missing values present.

Table 6.1: Benchmark data selected for classification from the UEA & UCR Time Series Classification Repository [Bag+17].

Name of Dataset	Samples	Classes	Data Origin
Beef	60	5	Spectrograph
ChlorineConcentration (ChlorineConc)	4307	3	Simulated
Earthquakes	461	2	Vibrations
ECG5000	5000	5	ECG measurement
FaceFour	112	4	Images
FordA	4921	2	Engine noise
LargeKitchenAppliances (L.K.Appliances)	750	3	Energy consumption
Meat	120	3	Spectrograph
OliveOil	60	4	Spectrograph
OSULeaf	442	6	Images
Symbols	1020	6	Images
SyntheticControl (SyntheticCont)	600	6	Simulated
Trace	200	4	Simulated
TwoLeadECG	1162	2	ECG measurement
Worms	258	5	Motion

Each of the data sets has been reviewed prior to application in terms of standardization, since incorrect or non-existent standardization can affect the results during classification [Bag+17]. Each of the selected data sets originally contains standardized samples of zero mean and unit standard deviation.

6.1.2 Classification methods

To generate classification results for fusion, the open source software package WEKA [Hal+09] is used to apply the classification methods. The different hyperparameters of each classifier are not considered, so the default-parametrization is used. The optimization of classifier inherent parameters is a highly problem specific task [Zhe15]. Different data characteristics usually require different parameter adjustments, so the application of optimized models on differing data sets would lead to highly biased results, when not specifically optimized for each of the given problems itself [Bag+17]. An optimization process requires additional amount of data, which is often unfeasible due to the already limited number of samples available. Therefore,

Table 6.2: Mean accuracy [%] for each classifier and data set.

Classifier Data set	C1	C2	C3	C4	C5	C6	C7	C8
Beef	58.33	**81.67**	46.67	53.33	61.67	**81.67**	25.00	33.33
ChlorineConc	92.80	**99.98**	61.07	41.30	99.33	70.20	38.26	37.98
Earthquakes	71.96	73.70	**79.78**	72.39	71.96	69.57	59.57	55.43
ECG5000	92.90	94.60	92.72	86.82	93.30	**94.80**	60.98	73.30
FaceFour	85.45	**96.36**	80.91	93.64	95.45	93.64	62.73	57.27
FordA	51.32	75.26	52.93	48.80	**68.27**	49.25	50.35	49.31
L.K.Appliances	54.67	36.27	45.73	50.27	**58.27**	45.87	44.80	39.33
Meat	95.00	**100.0**	95.00	97.50	99.17	**100.0**	55.00	83.33
OliveOil	88.33	**91.67**	80.00	**91.67**	88.33	88.33	78.33	71.67
OSULeaf	42.95	61.59	37.05	37.05	**65.23**	51.82	37.50	33.18
Symbols	94.41	**96.96**	90.00	93.63	96.37	95.88	67.16	65.10
SyntheticCont	80.00	94.50	81.00	**97.00**	90.33	95.50	56.33	61.33
Trace	89.50	88.50	68.50	87.50	**91.50**	78.00	41.50	41.50
TwoLeadECG	93.79	**99.66**	97.50	78.28	99.40	99.48	52.67	52.93
Worms	**47.06**	37.25	41.96	32.16	46.67	36.47	29.41	24.71

to bypass the problem of limited data, and as done in several other contributions ([Mor^{+}06], [AS06]), the applied classifiers are implemented with the default parameter setting applying the WEKA machine learning toolbox version 3.8.2 [Hal^{+}09]. The classification methods to be discussed are C4.5 Decision Tree (C1), Multilayer Perceptron (C2), Radial Basis Function Network (C3), Naive Bayes (C4), K-Nearest Neighbors (C5), Support Vector Machine (C6), Expectation Maximization (C7), and Simple K-Means (C8). To enable the application of unsupervised clustering methods with respect to the problem of classification, the WEKA ClassificationViaClustering procedure [Hal^{+}09] is implemented, which generates a mapping between the clusters derived from training data and their corresponding class labels in a supervised manner [Bou^{+}13]. The results of the individual classifiers (mean accuracy for all five folds) are given in Table 6.2. The best as well as the worst result for each considered data set is printed in bold or is underlined respectively.

6.1.3 Fusion algorithms

For a consistent distribution of fusion methods to all attributes, seven different fusion methods are selected. In Table 6.3, the methods Majority Voting (MV), Highest Rank (HR), Borda Count (BC), Bayesian Combination Rule (BCR), Behavior Knowledge Space (BKS), Logistic Regression (LR), and Fuzzy Integral (FI) are shown with their specific characteristics according to the utilized type of classifier output, the necessity of training parameters, and whether they are class-conscious (bold) or class-indifferent (not bold). In the following analysis (in accordance to [KBD01]) the classifier output levels are distinguished into single class label outputs (abstract) and ranking or measurement outputs (soft).

6.1.4 Procedure

Although several measures for assessing the performance of classification algorithms exist [Zhe15], the focus of the conducted numerical experiments is given to the

Table 6.3: Different fusion methods ordered by characteristics (class-conscious methods are printed in bold) (referring to [KBD01]).

		Trainable method	
		No	Yes
Output level	Abstract	**Majority Voting**	Bayesian Combination Rule Behavior Knowledge Space
	Soft	**Highest Rank** Borda Count	Logistic Regression **Fuzzy Integral**

analysis with respect to the overall classification accuracy obtained by classifier fusion. Therefore the accuracy gain is considered.

According to k-fold cross validation, data sets are divided into k folds. Random partitioning of data into k disjoint folds is done [Mor+06, Krs+14]. The partitioning conducted within each of the outer loops of cross validation is the same for every classifier and fusion method (due to implementing seeded random partitioning in WEKA) [Bou+13]. This way, by generating the exact same folds of data for every classifier and fusion method, all of the conducted experiments are completely reproducible when based upon the attached basic partitions of data. To prevent additional bias caused by a dissimilar deployment of classes between training and test data, each of the derived partitions possesses the same class distribution as the corresponding original set of samples by implementing stratification of partitions [BF04, Bou+13]. The suggested number of folds using cross validation according to investigations in [Koh95] and [RPL10], strongly depends on the stability of applied induction algorithms. For larger values ($k = 10 - 20$), in [Koh95] a reduced variance by simultaneous increasing bias of estimates is noted, while for smaller values ($k = 2$) the variance increases significantly. In a similar way, the recommendations proposed by [RPL10] ranges between two to ten folds. Here the applied number of folds is consequently defined as $k = 5$ for both inner and outer loop of cross validation, providing a trade-off between bias and variance while restricting computational effort. The considered data set is divided into 5 folds, four of these folds are used to train, one fold is used to test the classifiers. Using the test fold, also the fusion methods are applied to compare the fusion performance with the individual classifier performance. This training/test procedure is repeated five times. According to [VS06], the resulting mean value (here of gain of accuracy) is used to compare the performances.

As explained, some of the selected fusion methods need also an additional training prior to the fusion process. Following the contributions [Krs+14, VS06, Won15] generating unbiased performance estimates, every aspect of parameter tuning or selection should be included within the procedure of cross validation itself, so a nested approach of cross validation is considered. This means, that every training data set (consisting of four folds of the entire data set) is again divided into five folds. Four of these folds are used for training of the classifiers and the 5th fold is used to calculate the necessary parameters for the fusion method. This is also done five times for each training data set, so 25 times in total. From all calculated values, the mean value is set as the final parameter used in the fusion process.

6.2 Numerical analysis

To evaluate the performance improvement and the influences of different characteristics, the results are discussed in the following. First the overall performance will be shown. The results related to the characteristics of fusion methods and data sets are discussed subsequently.

6.2.1 Mean performance of fusion methods

To analyze the overall fusion performance, the accuracy gain calculated for each fusion method and data set is shown in Table 6.4. The best as well as the worst result for each considered data set is printed in bold or is underlined respectively, not negative values are highlighted in green. For all 105 combinations of data set and fusion method, only 13 times no deterioration compared to the best individual classifier occur. Considering the number of best and worst results, LR is the fusion method with the best performance (12 times best result out of 15 data sets). This is also confirmed calculating the mean percentage (-0.95 %) over all considered data sets (Table 6.4 last row). The second best results are produced using FI (mean percentage = -3.14 %). The mean percentage for the methods BCR, BC, and MV

Table 6.4: Mean accuracy gain [%] for each fusion method and data set [RKS19].

Fusion method Data set	MV	HR	BC	BCR	BKS	LR	FI
Beef	-18.33	-38.33	-23.33	-25.00	-48.33	**-5.00**	-11.67
ChlorineConc	-16.39	-52.52	-15.46	-0.02	-5.27	**0.00**	-0.14
Earthquakes	-2.59	-24.07	-1.73	**0.00**	-1.95	-0.43	-4.76
ECG5000	-0.66	-25.30	-0.76	-1.68	-2.78	**0.08**	-0.52
FaceFour	**-1.78**	-32.02	-2.69	-10.63	-26.68	**-1.78**	-2.69
FordA	-14.83	-24.36	-15.14	-2.48	-3.35	**-0.06**	-6.38
L.K.Appliances	-8.13	-22.27	-5.73	-4.80	-14.13	**0.93**	-2.40
Meat	**0.00**	-22.50	**0.00**	-7.50	-10.00	**0.00**	-5.83
OliveOil	-5.00	-10.00	-5.00	-11.67	-16.67	**-3.33**	-5.00
OSULeaf	-15.63	-28.95	-12.89	-4.08	-35.28	**0.00**	-3.42
Symbols	-1.27	-12.06	-2.25	-3.24	-19.31	**0.39**	-0.78
SyntheticCont	0.17	-30.67	-1.33	0.00	-17.17	**1.67**	1.00
Trace	-9.00	-32.00	-13.00	-3.50	-19.00	-5.00	**-2.50**
TwoLeadECG	-1.38	-31.24	-1.63	-0.17	-0.52	-0.26	**-0.09**
Worms	-9.28	-20.94	-2.71	-5.81	-25.13	**-1.52**	-1.95
Mean percentage	-6.94	-27.15	-6.91	-5.37	-16.37	**-0.95**	-3.14

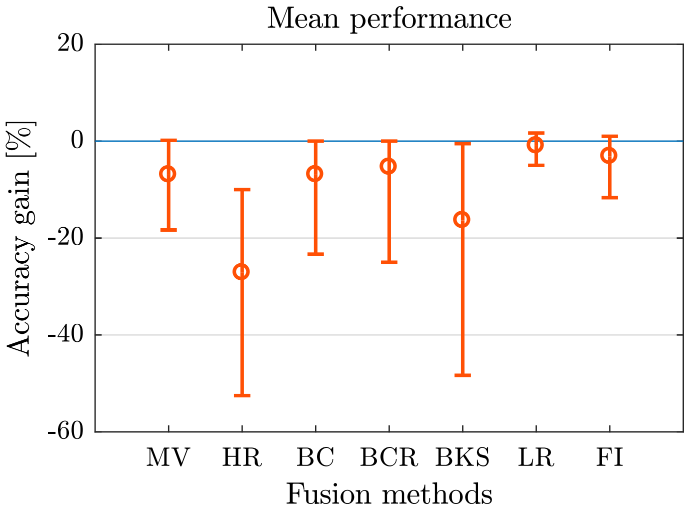

Figure 6.1: Mean, minimum and maximum accuracy gain over all data sets for each fusion method [RKS19].

are in a similar range (−5.37 % to −6.94 %). The fusion method HR shows the worst results for the considered data sets, for 10 out of the 15 data sets as well as for the mean percentage, HR produces the least accuracy gain. The BKS shows worst results for the remaining 5 data sets. In Figure 6.1 the mean percentage as well as minimum and maximum value of accuracy gain are plotted for different fusion methods.

The results show, that also the range between minimum and maximum value is lower for those fusion methods with better mean value (LR and FI) and higher for those with worse results (BKS and HR). Although there are strong tendencies for specific fusion methods, the variety of results is very high, not only for the different fusion methods, also for different data sets. The dependency of the results to the different characteristics of fusion methods and data sets is analyzed in the following.

Lessons learned In most cases fusion lead to a deterioration in accuracy compared to the best individual classifier. The LR fusion method leads to the best, HR to worst results regarding the considered data sets and fusion methods.

6.2.2 Performance related to fusion method characteristics

The distinction between trainable and non trainable, class-conscious and class-indifferent, as well as abstract and soft level-supported methods is illustrated in Figure 6.2.

Considering the type of classifier output, which is used for the fusion methods, it can not be stated that one type outperforms the other one (see Figure 6.2 top left).

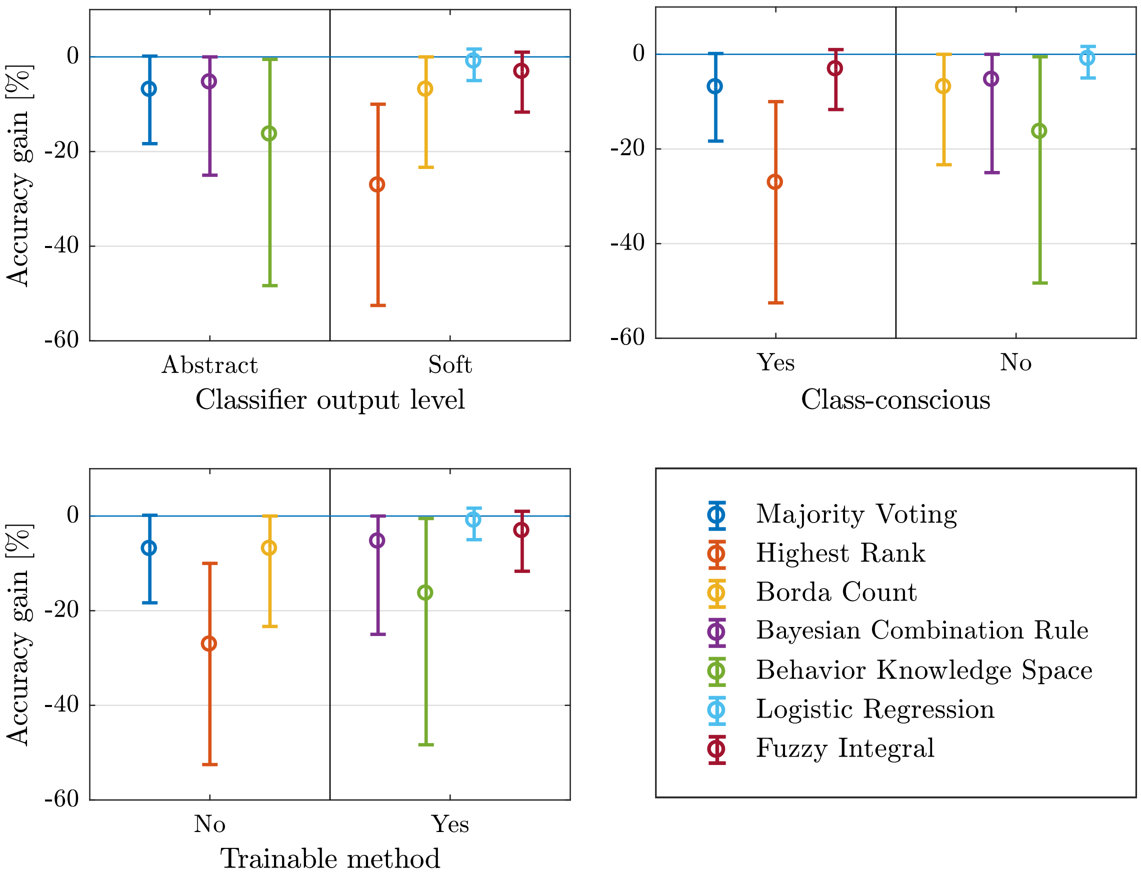

Figure 6.2: Mean, maximum and minimum accuracy gain over different characteristics of fusion methods [RKS19].

Fusion methods with best and worst results (LR and HR) both use soft classifier outputs. Further the mean value of the individual mean values (from Table 6.4) is similar in both categories (abstract: −9.56 %, soft: −9.53 %). For the considered data sets and fusion methods, the type of classifier output has no significant influence although the information content is higher using the soft level.

Beside the mentioned type of classifier outputs, also the influence of different use of these outputs is considered. In Figure 6.2 top right the results are distinguished in class-conscious and class-indifferent fusion methods. The results show, that the two best performances produced by LR and FI, as well as the two worst results produced by BKS and HR are in different categories, whereas the best performance is produced by a class-indifferent method, the worst by a class-conscious one. The mean of the individual mean values also show a difference (class-conscious: −12.41 %, class-indifferent: −7.40 %) and a tendency to class-indifferent methods.

Taking the necessity of an additional training into account, the best three performance values can be reached using trainable fusion methods. The mean of individual mean values show a smaller absolute value for trainable methods (not trainable: −13.67 %, trainable: −6.47 %). This result shows, that additional training, providing additional information used in the fusion process, leads to better results.

Lessons learned The level of classifier output does not have significant influence on the fusion performance. Class-indifferent and trainable methods perform slightly better than class-conscious and non trainable methods.

6.2.3 Performance related to data characteristics

The following paragraphs focus the impact of data inherent characteristics. Therefore the mean, minimum and maximum values of accuracy gain are plotted for each fusion method and data set separately over the considered data properties.

Number of classes

The first data characteristic analyzed is the number of classes. In Figure 6.3 each plot contains the accuracy gain for the 15 data sets depending on their number of classes of one fusion method. The results show, that the number of classes has not an influence on every fusion method. The methods MV, HR, BC, and FI do not show a significant change in mean value, only the range between maximum and minimum value increases and reaches its maximum at 5 classes considering the methods HR and BC. Utilizing the fusion methods BCR, BKS, and LR, a decrease of accuracy gain as well as an increase of the range can be observed for increasing number of classes. For the LR method, the occurrence of this tendency is not as significant as for the BCR and BKS fusion methods. Both methods (Bayesian Combination Rule and Behavior Knowledge Space) rely on the previous training of method inherent parameters. In the case of BCR, the probability matrix which has to be computed prior to classification, grows quadratic with the number of classes. Thus an increasing number of classes requires a higher number of training samples. In a similar way the application of Behavior Knowledge Space relies on the previous computation of probabilities for each of the possible combinations of labels generated by the different base classifiers. Hence the knowledge space also grows exponentially with the number of possible classes, which in the case of restricted data for training also impedes derivation of proper results.

Lessons learned For most fusion methods an increasing number of classes (up to 5 classes) lead to a decreasing performance. For a small number of classes, BCR, BKS, LR, and FI are suitable, but for a higher number of classes, only LR and FI are recommended.

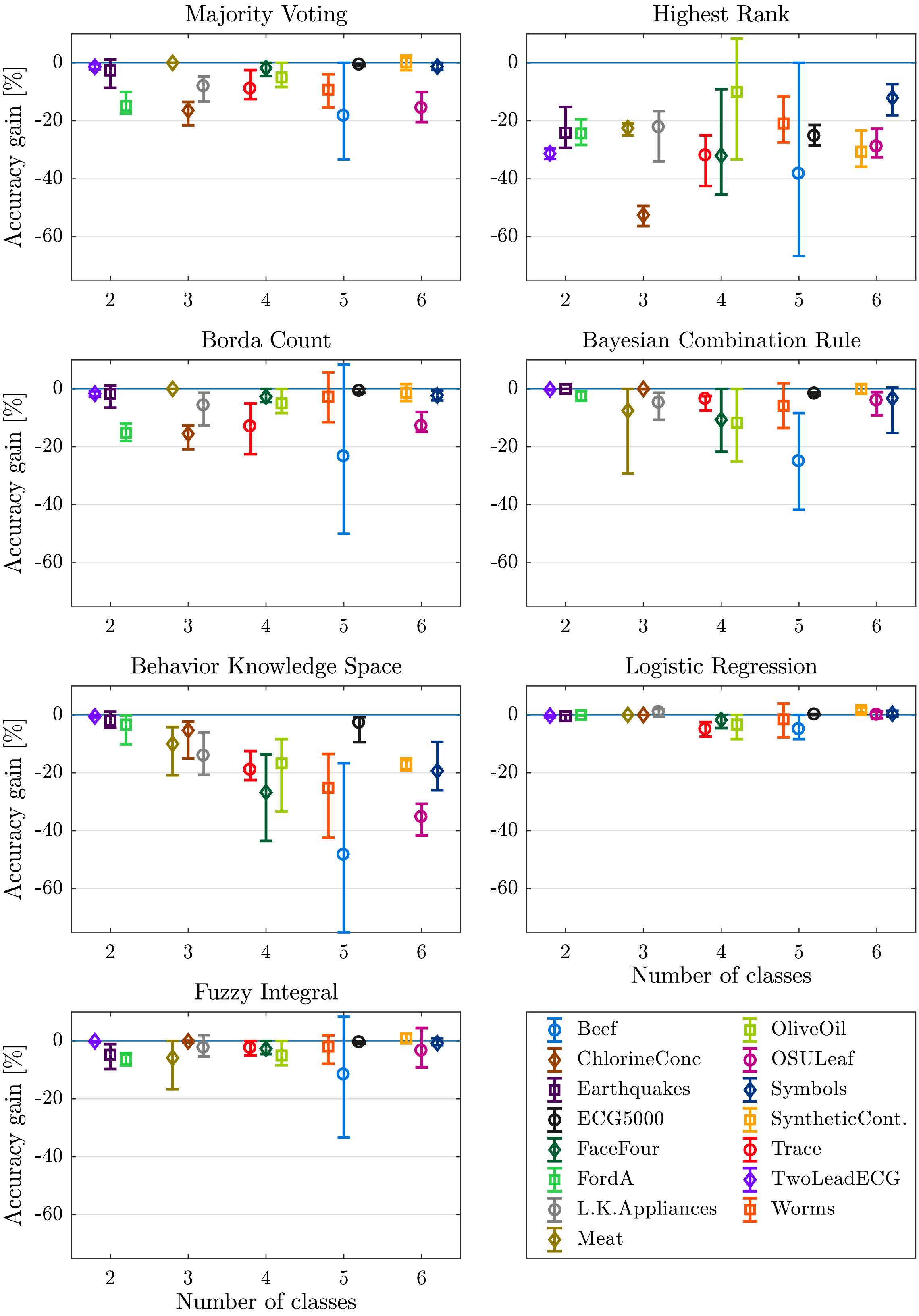

Figure 6.3: Mean, minimum and maximum gain in accuracy over number of classes, with respect to each data set and fusion method considered [RKS19].

Number of samples

In Figure 6.4 the accuracy gain is plotted over the number of samples provided by the individual data set. Here the number of samples of the applied data sets ranges from 60 samples (Beef and OliveOil) to 5000 samples (ECG5000). The x-axis is scaled logarithmic because of a slight majority of lower numbers. The mean values of accuracy gain reached with the fusion methods MV, HR, LR, and FI show no clear tendency for increasing number of samples except of the data set ChlorineConc. (4307 samples) using HR method. Considering the methods BC, BCR, and BKS the loss of accuracy decreases for increasing number of samples. Regarding the range between minimum and maximum accuracy gain, the range decreases for all methods except of LR for increasing number of samples. Some exceptions for some data set/fusion method combinations should be stated: The data sets FaceFour (112 samples) and Meat (120 samples) show a small range and also a higher mean value using the fusion methods MV and BC, also showing a small number of samples. All exceptions only appear for fusion methods, where no additional training of fusion parameters is necessary. The most drastic impact can be noticed for the methods of BCR as well as BKS, with an improvement of over 20 % and 40 % respectively.

Lessons learned All trainable methods show increasing performance for increasing number of samples. If a large amount of data is available, trainable methods should be preferred, if only a small number of samples can be used, LR or FI should be applied.

Entropy of classes

To evaluate the dependency of fusion results on the evenness of class distribution, the accuracy gain is plotted against the entropy of classes for each data set and fusion method in Figure 6.5. The within this work classified sets of data comprise a range of entropy between 0.7245 to 2.5850 Bit for the Earthquakes and SyntheticControl data set respectively. Regarding the methods BCR, BKS, and LR the results show a more or less significant influence on the mean value of accuracy gain. With increasing entropy, the mean value decreases except of data sets with entropy of more than 2.5 Bit (OSULeaf, Symbols, and SyntheticControl). The most clearly observable tendency can be recognized for the method of BKS. The methods MV, HR, BC, and FI do not show a tendency of mean value for changing entropy of classes. The range between minimum and maximum value increases with increasing entropy. This can be observed for all methods, the most drastic impact on maximum-minimum range can be noticed for the methods of BKS, BCR, BC, and HR. The entropy of classes, as a measure of class distribution, reaches the highest value with respect to a specific number of possible classes, if all of the possible labels are represented by the

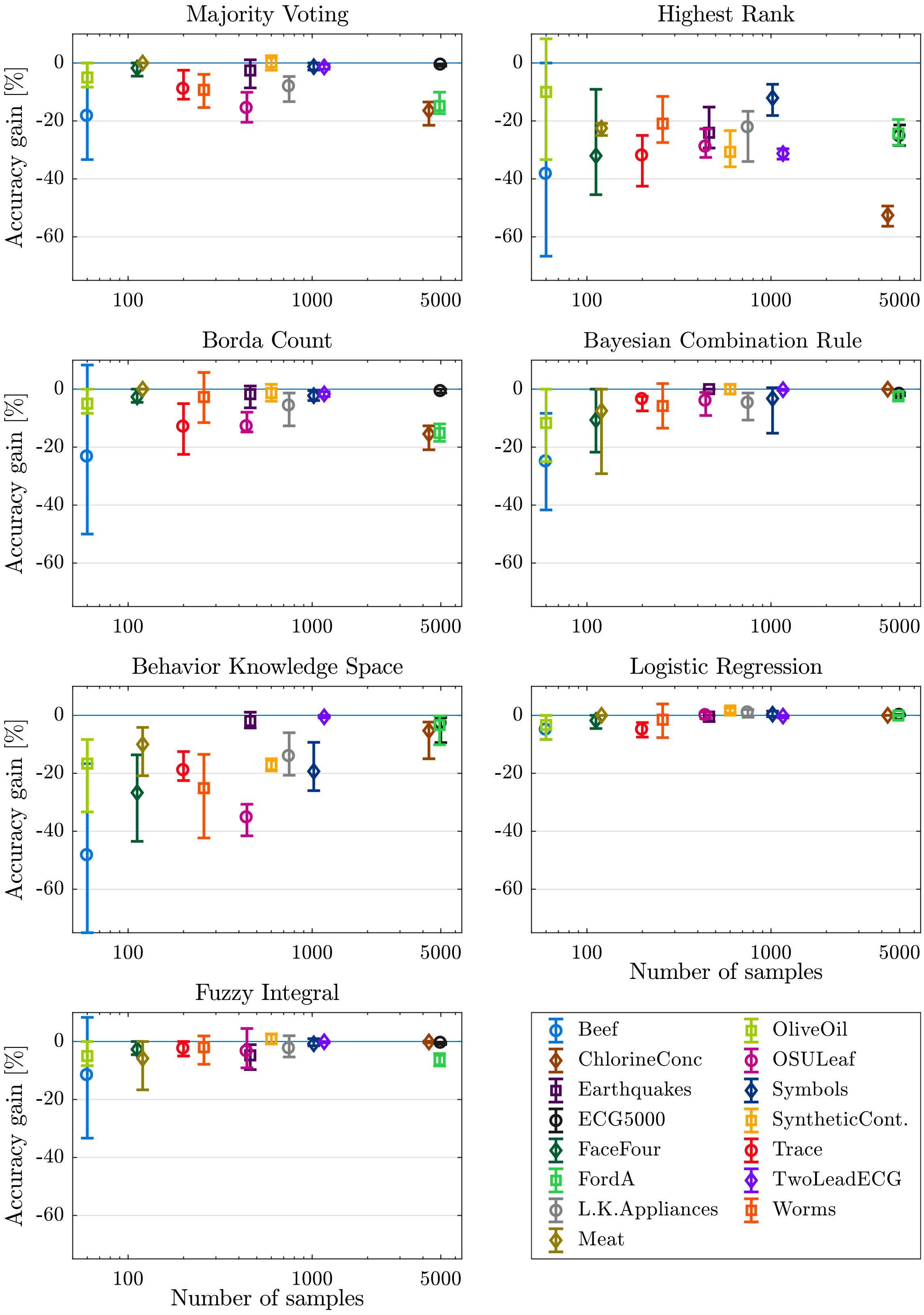

Figure 6.4: Mean, minimum and maximum gain in accuracy over number of samples, with respect to each data set and fusion method considered [RKS19].

same amount of samples within the considered set of data. A higher value of entropy corresponds to a higher probability for the samples of a specific class, to take part in the set applied for inducing the models for classification. The fact that skewed sets of data in many cases tend to cause overly optimistic results in terms of the resulting accuracy, may explain the observed behavior for certain fusion methods.

Lessons learned An increasing entropy of classes lead to a decreased performance for most fusion methods, although the information content is higher and the classes are more even distributed for high entropy of classes. For a high entropy of classes, only LR and FI show good performance, while only HR is not recommended to be implemented for data with small entropy of classes.

Entropy of attributes

Not only the entropy of classes, also the entropy of attributes is considered in this analysis. As mentioned, for each attribute of one data set, one entropy value is calculated. To get one value specific for one data set, the mean value of the entropy of all attributes is calculated. The within this work selected sets of benchmark data, comprise a range of entropy from 1.1211 Bit to 8.0988 Bit for the Beef and FordA data set respectively. The mean, maximum and minimum of accuracy gain in dependency of the entropy of attributes is shown in Figure 6.6 for each data set and fusion method. Considering the mean value of accuracy gain, a small increase in mean for increasing entropy of attributed can be observed using the methods MV, BC, and BCR, whereas using the BKS fusion method, the influence is significant. The other methods (HR, LR, and FI) show evenly distributed mean values for all entropies. The data sets ChlorineConc and FordA show the highest entropy of attributes, but the mean of accuracy gain decreases significantly for one or both of these data sets when using the fusion methods MV, HR, or BC (again only the not trainable methods). For all fusion methods the range between minimum and maximum decreases significantly for increasing entropy of attributes. The mean entropy of attributes is defined as the arithmetic mean over all single entropy of attributes. Given the fact that an attribute with an corresponding low value of entropy comprises less change in magnitudes, these attributes tend to provide only a slight amount of additional information usable for the task of classification [MST94]. The in Figure 6.6 illustrated results support these quotations, since for lower values of mean entropy, and therefore less attribute inherent information, the performance of the applied methods of decision fusion is clearly impaired.

Lessons learned An increasing performance can be observed for increasing entropy of attributes, because more information can be extracted using data sets with a high entropy of attributes. Considering data with low entropy of attributes, only

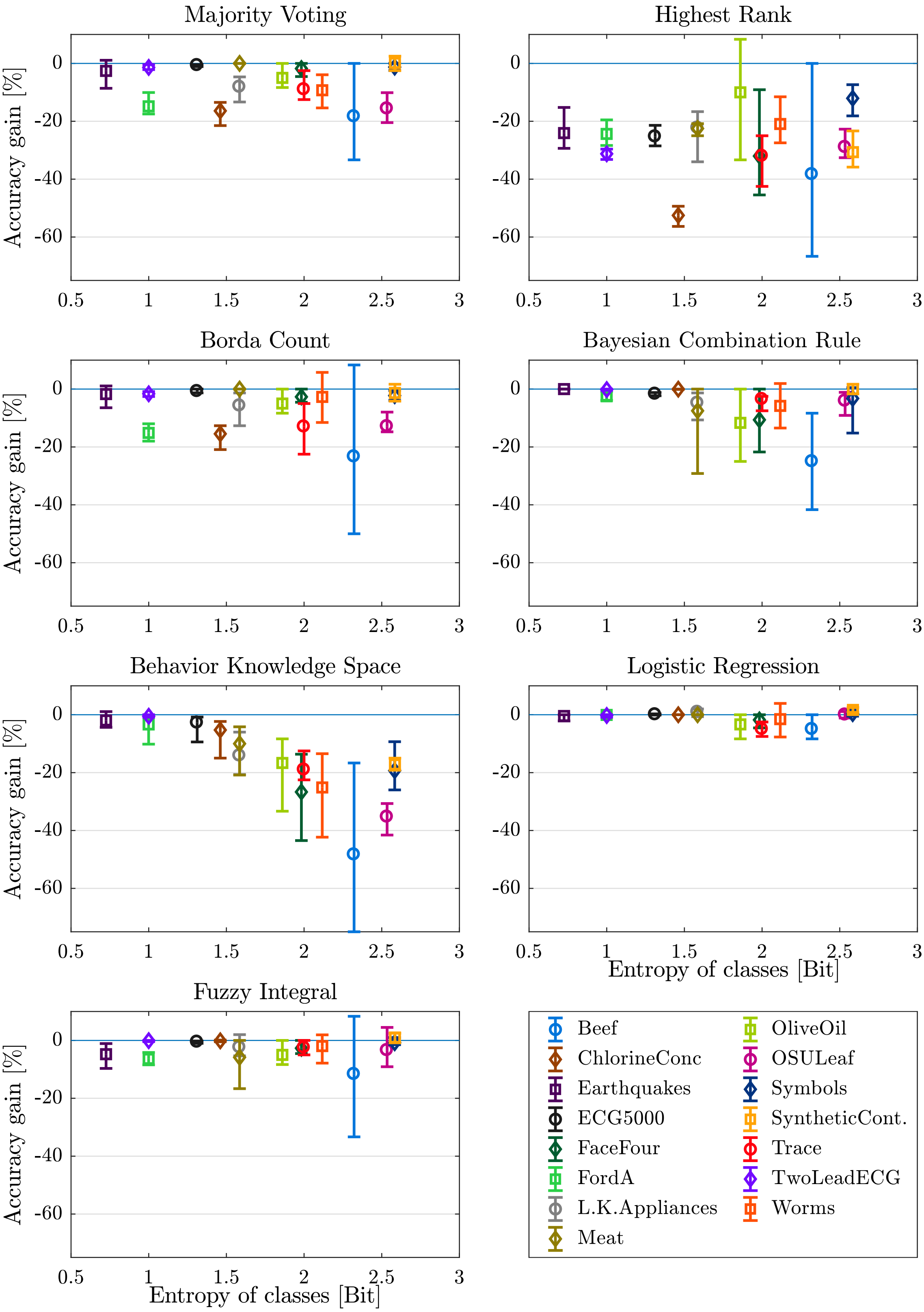

Figure 6.5: Mean, minimum and maximum gain in accuracy over entropy of classes, with respect to each data set and fusion method considered [RKS19].

LR should be used as fusion method. For data with higher entropy of attributes, LR and also BCR, BKS, and FI can be recommended.

6.3 Conclusions from the analysis of data and fusion method characteristics

The question if fusion methods are increasing the overall performance and which characteristics are influencing the fused performance is considered. A summary of all lessons, which can be generated from the numerical analysis is listed in Table 6.5.

The overall fusion performance illustrates the differences between the considered fusion methods. While the Logistic Regression outperforms the other fusion methods, the performance of the best individual base classifier can only be exceeded in 5 out of 105 cases. In 92 of 105 cases, using fusion lead to a deterioration of performance compared to the best individual classifier performance.

The type of classifier output (abstract or soft) as well as how these outputs are used (class-conscious or class-indifferent) have no significant influence on the fusion performance, whereas the trainable methods show slightly better performance than methods without the additional training prior to the fusion process.

Considering data characteristics, the results show, that a higher number of classes leads to worse performance for some of the fusion methods. The methods LR and FI show the most constant performance for all numbers of classes, while BKS shows the highest sensitivity to increasing class number.

The more samples the data set have, the more information can be used for training the base classifiers as well as the parameters of fusion methods if necessary. This results in a better performance of the trainable methods for data sets with higher number of samples and can be concluded from results by the increasing mean value only for trainable methods (BCR, BKS, LR, and FI).

Although a higher entropy of classes denotes more even distributed classes in the considered data set, a decrease in performance (mean and also range between maximum and minimum accuracy gain) can be observed for most of the fusion methods. Considering data sets with low entropy, all fusion methods except of HR (fusion method with worst overall performance) show similar and good performance. Increasing the entropy of classes, only using LR and FI are recommended to reach good performance.

Complementary to the entropy of classes, the higher entropy of attributes lead to better performance of fusion methods. The higher the entropy, the more information can be extracted from the attributes, which can also be concluded from the results of all fusion methods.

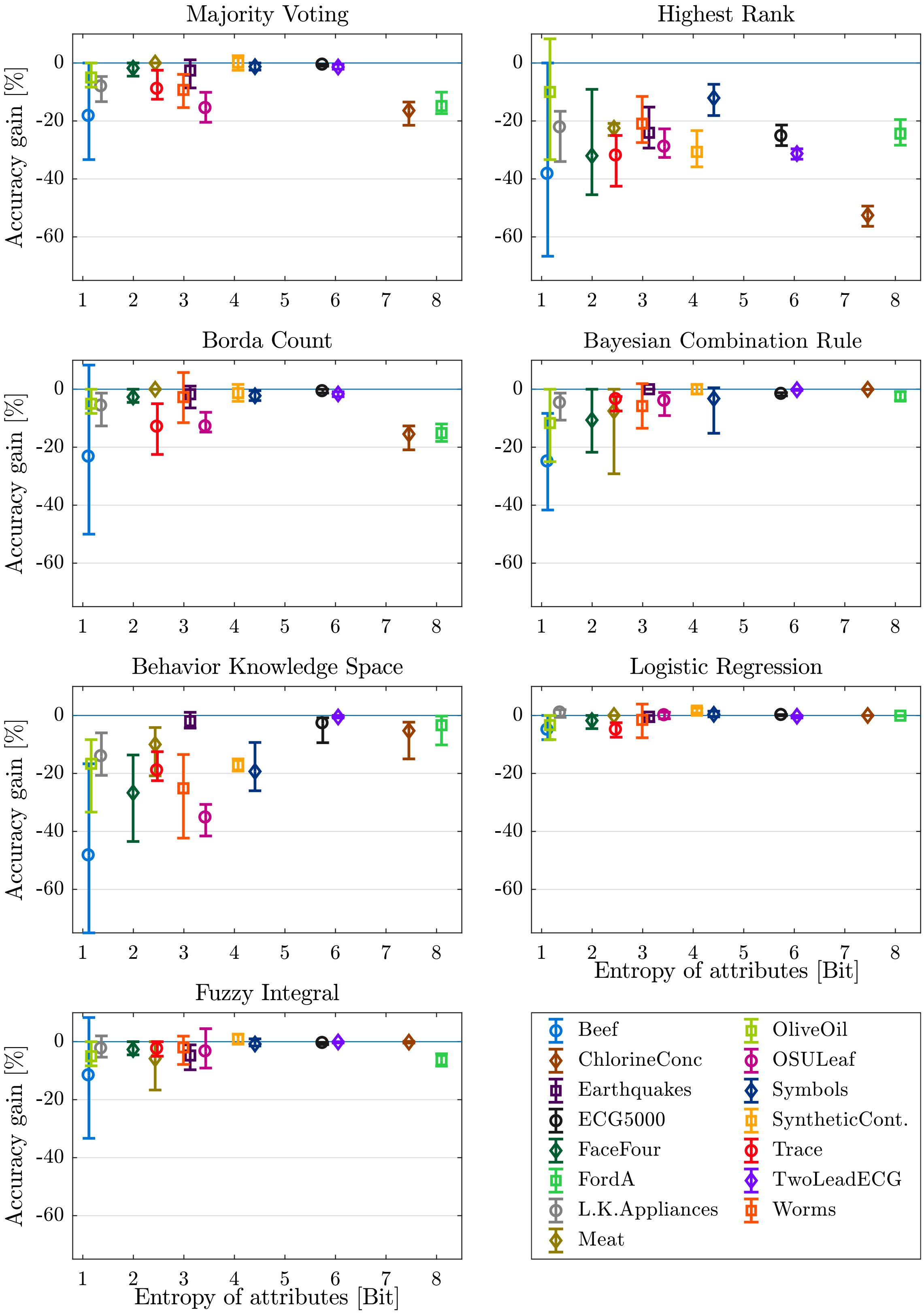

Figure 6.6: Mean, minimum and maximum gain in accuracy over mean entropy of attributes, with respect to each data set and fusion method considered [RKS19].

Considering all data characteristics, MV, LR, and FI show the least sensitivity to a change in these characteristics. While LR is performing best in the overall performance, the absolute influence of the changes is the least using LR. Hence the LR method is denoted as the least sensitive method. The fusion method BKS is most sensitive to changes of applied data characteristics.

Table 6.5: Lessons learned from the numerical analysis [RKS19].

Characteristic	Lessons learned
Overall performance	In most cases fusion lead to a deterioration in accuracy compared to the best individual classifier. The LR fusion method leads to the best, HR to worst results regarding the considered data sets and fusion methods.
Classifier output level	The level of classifier output does not have significant influence on the fusion performance.
Use of classifier output	Class-indifferent methods perform slightly better than class-conscious methods.
Necessity of training	Trainable methods perform slightly better than non trainable methods.
Number of classes	For most fusion methods an increasing number of classes (up to 5 classes) lead to a decreasing performance. For a small number of classes, BCR, BKS, LR, and FI are suitable, but for a higher number of classes, only LR and FI are recommended.
Number of samples	All trainable methods show increasing performance for increasing number of samples. If a large amount of data is available, trainable methods should be preferred, if only a small number of samples can be used, LR or FI should be applied.
Entropy of classes	An increasing entropy of classes lead to a decreased performance for most fusion methods, although the information content is higher and the classes are more even distributed for high entropy of classes. For a small entropy of classes, only LR and FI show good performance, while only HR is not recommended to be implemented for data with higher entropy of classes.
Entropy of attributes	An increasing performance can be observed for increasing entropy of attributes, because more information can be extracted using data sets with a high entropy of attributes. Considering data with low entropy of attributes, only LR should be used as fusion method. For data with higher entropy of attributes, LR and also BCR, BKS, and FI can be recommended.
Sensitivity to all data characteristics	Using the fusion method LR, the results are at least sensitive to the changes in the data characteristics, while BKS shows the most sensitivity.

7 Summary, conclusion, and future work

7.1 Summary and conclusion

The research work presented in this thesis is related to find influencing factors affecting the fusion performance. The literature review of this thesis emphasizes, that fusion results are depending on used application data, the selection strategy, or the fusion method itself. The open question is to define the characteristics of data, classifier, or fusion method leading to an influence on the performance. Therefore in this thesis, each part of the multiple classifier system is considered and analyzed to identify, which data characteristics, classifier properties, selection strategies, and fusion method characteristics lead to a reliable information fusion and in which cases fusion is not worth to be established.

To examine the effects from data structure to resulting accuracy using different ensemble selection strategies and fusion methods two strategies for ensemble selection are used: SCE and DCE. Seven data sets from the UCR Time Series Classification Repository are classified with eleven classifiers using WEKA. As competence measurements, the overall accuracy of the individual classifier and the precision of each class for each classifier is used respectively. Two experimental applications for condition monitoring (fault diagnosis of hot rolling mill and damage detection in composites) are used to further analyze the selection strategies. All data sets are divided into validation and test data using cross-validation. The results from the numerical analysis show, that the applied SCE and DCE strategies can improve the overall performance. The analysis using SCE generates better results according to the fused accuracy than DCE. An optimal number of classifiers selected for an ensemble can not generally be given. The optimal number is depending on the total number of available classifiers. In the evaluated examples, the number of classifiers selected in the ensemble generating the best results is around half of the number of available classifiers. In case of eleven classifiers, the selection of five classifiers produces better results than the selection of two classifiers. For fault diagnosis of hot rolling mill, six base classifiers are available. The best result can be achieved using four classifiers for the selected ensemble. The experimental results of damage detection in composites, where four base classifiers are available, the best result is achieved using two classifiers. During analysis using all combinations of selected classifiers, the best results are not always generated using the best base classifiers. In some cases, the selection of best in combination with worse classifiers, higher accuracies can be achieved.

To analyze the dependency of the precision values of one classifier to the overall fused accuracy and to get requirements for a good fusion performance a fictional classifier is introduced. Four benchmark data sets with two classes and one experimental data set with four classes (fault diagnosis of hot rolling mill) are considered and fused in different ensembles to show the dependencies of the precision values on overall fusion accuracy. To ensure a certain kind of generalizability all results are based on the 3-fold cross-validation. From the results it can be concluded, that varying the precision values an improvement of fused accuracy using the FC compared to the accuracy of fusion without FC can be obtained. The related precision values vary for every case, so no general information about precision values can be given, nevertheless for each data set and ensemble individually the improvement potentials can be evaluated. The analysis of only one additional classifier revealed that an additional classifier, which is worse than the other one, can have a wide range of precision values leading to the same fused accuracy. The best individual one should have specific precision values when combining with an additional worse classifier. The dependency on used ensembles, number of available data samples, and data sets can also be concluded from the results, because the improvement potentials vary for different combinations. In general and by the nature of the problem no overall suggestion for the requirements for good fusion performance can be established. To overcome this problem it can be shown that the considered factors (precision values) have an influence on the performance. Furthermore the potential to improve the fusion performance using a fictional classifier can be shown. If an improvement is possible, the precision values as well as the ensemble selection and classifier set as FC from training can be used for fusion of new samples. In general, the evaluation using the introduced concept of a fictional classifier can help to analyze the improvement potentials. It can assist by deciding if a fusion of results will be successful or the individual classifier should be preferred.

To answer the question, which influence the performance measure of a classifier used for fusion has on the fusion performance, a new concept of POD-based fusion is introduced. The concept is applied to fault diagnosis in elastic structures with an elastic beam as example. The measurements used are acceleration, strain, and displacement. The POD characterization depends on the sensor position, fault position, and the feature selected. Using the fusion approach introduced, the probability of the existence of a fault can be obtained based on the individual performance and assignments of the sensor-/feature-based statement. The use of probability estimations as classifier output is also possible using the BCR and WV fusion methods. The results show a crisper decision for the fused results than for individual classifiers.

To investigate relationships between fusion methods and data characteristics, the question to be answered is: Which fusion method improves in which case the overall performance? Therefore 15 different sets of benchmark data with different number of classes and samples as well as different entropy of classes and attributes were classified by eight base classifiers implemented using the WEKA machine learning toolbox. The generated classification results were fused by seven selected fusion algorithms. The fusion methods use different types of classifier outputs (abstract and soft) in different ways (class-conscious or class-indifferent). Some need additional training prior to fusion, some not. During experiment, nested 5-fold cross validation is used to distribute the data sets into training and test for classifiers, and training and validation for fusion methods to obtain representative and (with the given restrictions) generalized results. The main and most important result is that in most cases the use of fusion methods do not outperform the maximum individual classifier performance. However, the use of fusion has advantages like insensitivity to overfitting or redundancy. The numerical analysis clearly points out, that trainable methods produce better results, while Logistic Regression outperforms the other fusion methods. Considering the data characteristics, a lower number of classes, more available samples (especially for trainable fusion methods), a lower entropy of classes, and a higher entropy of attributes lead to better performance for most of the used fusion methods. The results lead to the conclusion, that the fusion performance strongly depends on the individual fusion method in combination with data characteristics.

The results and analyses in this thesis lead to the conclusion, that the fusion performance is depending on characteristics of each part of the multiple classifier system. Several ways to identify these influencing factors and recommendations to ensure a reliable fusion performance are given.

7.2 Future work

In this thesis the accuracy as well as the precision values are used for selection of the classifier ensemble. Other measures like the recall and the false alarm rate, as well as diversity measures can be considered and analyzed to show the influence of the performance measure on fusion performance during selection of ensembles.

The concept of the POD-based fusion is introduced and applied to one example to evaluate the feasibility of the introduced concept. To examine the influence on the fusion performance, further experiments have to be done where all sensor/feature combinations are detecting a fault or not. Using the results of the experiments, individual accuracy as well as fused accuracy can be compared.

Using the results of this thesis, an approach selecting the ensemble, the performance measures of the classifier, and the fusion method depending on the data set can be realized.

Bibliography

[And+10] Anderson, D. T.; Bezdek, J.C.; Popescu, M.; Keller, J.M.: Comparing fuzzy, probabilistic, and possibilistic partitions. In: *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 5, pp. 906-918, 2010.

[Ann17] Annis, C.: *Statistical best-practices for building Probability of Detection (POD) models*, R package mh1823, http://StatisticalEngineering.com/mh1823/, 2017.

[ARS19] Ameyaw, D. A.; Rothe, S.; Söffker, D.: A novel feature-based Probability of Detection (POD) assessment and fusion approach for reliability evaluation of vibration-based SHM systems. In: *Structural Health Monitoring*, 2019, accepted.

[ARS18a] Ameyaw, D. A.; Rothe, S.; Söffker, D.: Adaptation and implementation of Probability of Detection (POD)-based fault diagnosis in elastic structures through vibration-based SHM approach. *The 9th European Workshop on Structural Health Monitoring (EWSHM)* , Manchester, July 10-13, 2018.

[ARS18b] Ameyaw, D. A.; Rothe, S.; Söffker, D.: Probability of Detection (POD)-oriented view to fault diagnosis for reliability assessment of FDI approaches. In: *ASME 2018 International Design Engineering Technical Conferences & Computers, 30th Conference on Mechanical Vibration and Noise*, Quebec City, August 26-29, vol. DETC2018-85554, pp.V008T10A041, 2018.

[ARS18c] Ameyaw, D. A.; Rothe, S.; Söffker, D.: Fault diagnosis using Probability of Detection (POD)-based sensor/information fusion for vibration-based analysis of elastic structures. PAMM-Wiley Online, 2018.

[AS06] Ali, S.; Smith, K.A.: On learning algorithm selection for classification. In: *Applied Soft Computing*, vol. 6, no. 2, pp. 119-138, 2006.

[ASS14] Al-Shrouf, L.; Saadawia, M.S.; Söffker, D.: Improved process monitoring and supervision based on a reliable multi-stage feature-based pattern recognition technique. In: *Information Sciences*, vol. 259, pp. 282-294, 2014.

[Bag+17] Bagnall, A.; Lines, J.; Bostrom, A.; Large, J.; Keogh, E.: The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. In: *Data Mining and Knowledge Discovery*, vol. 31, no. 3, pp. 606-660, 2017.

[Bez+06] Bezdek, J.C.; Keller, J.; Krisnapuram, R.; Pal, N.: *Fuzzy models and algorithms for pattern recognition and image processing*, Springer Science & Business Media, New York, vol. 4, 2006.

[BF04] Bouckaert, R.R.; Frank, E.: Evaluating the replicability of significance tests for comparing learning algorithms. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, pp. 3-12, 2004.

[Bil15] Bilski, P.: Analysis of the classifier fusion efficiency in the diagnostics of the accelerometer. In: *Measurement*, vol. 67, pp. 116-125, 2015.

[Bou+13] Bouckaert, R.; Frank, E.; Hall, M.; Kirkby, R.; Reutemann, P.; Seewald, A.; Scuse, D.: Waikato Environment for Knowledge Analysis (WEKA) Manual for Version 3-7-8 (accessed: 2018-05-20), The University of Waikato, Hamilton, New Zealand, 2013.

[BSO14] Britto, A.S.; Sabourin, R.; Oliveira, L.E.S.: Dynamic selection of classifiers - a comprehensive review. In: *Pattern Recognition*, vol. 47, no. 11, pp. 3665-3680, 2014.

[Cai+10] Cai, Y.; Chow, M.-Y.; Lu, W.; Li, L.: Evaluation of distribution fault diagnosis algorithms using ROC curves. In: *IEEE PES General Meeting*, July 25-29, pp. 1-6, 2010.

[CFM09] Cobb, A.C.; Fisher, J.; Michaels, J.: Model-assisted probability of detection for ultrasonic structural health monitoring. In: *4th European-American Workshop on Reliability of NDE*, Berlin, Germany, June, 2009.

[Che+14] Chen, Y.; Keogh, E.; Hu, B.; Begum, N.; Bagnall, A.; Mueen, A.; Batista, G.: The UCR Time Series Classification Archive. [Online] Available: http://www.cs.ucr.edu/eamonn/time_series_data/, 2014 (accessed February 2014).

[CK95] Cho, S.-B.; Kim, J.H.: Combining multiple neural networks by fuzzy integral and robust classification. In: *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 25, pp. 380-384, 1995.

[Cri+15] Crivelli, D.; Guagliano, M.; Eaton, M.; Pearson, M.; Al-Jumaili, S.; Holford, K.; Pullin, R.: Localisation and identification of fatigue matrix cracking and delamination in a carbon fibre panel by acoustic emission. In: *Composites Part B: Engineering*, vol. 74, pp. 1-12, 2015.

[CSC18] Cruz, R.M.O; Sabourin, R.; Cavalcanti, G.D.C.: Dynamic classifier selection: Recent advances and perspectives. In: *Information Fusion*, vol. 41, pp. 195-216, 2018.

[Dem67] Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. In: *The Annals of Mathematical Statistics*, vol. 38, pp. 325-339, 1967.

[Dep09] Department of Defense: *Department of Defense Handbook, Nondestructive Evaluation System Reliability Assessment, MIL-HDBK-1823*. Department of Defense, 2009.

[DT98] Duin, R.P.W.; Tax, D.M.J.: Classifier conditional posterior probabilities. In: Amin, A.; Dori, D.; Pudil, P.; Freeman, H. (Eds.), *Advances in Pattern Recognition. SSPR/SPR 1998. Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, vol. 1451, pp. 611-619, 1998.

[Dui02] Duin, R.P.W.: The combining classifier: to train or not to train?. *Object recognition supported by user interaction for service robots*, Quebec City, Canada, August 11-15, vol. 2, pp. 765-770, 2002.

[EFB12] Eckstein, B.; Fritzen, C.; Bach, M.: Considerations on the reliability of guided ultrasonic wave-based SHM systems for CFRP aerospace structures. In: *Proceedings of the 6th European Workshop on Structural Health Monitoring*, Dresden, Germany, July 2-6, pp. 1-8, 2012.

[FCX15] Lustosa Filho, J.A.S.; Canuto, A.M.P.; Xavier-Junior, J.C.: An analysis of diversity measures for the dynamic design of ensemble of classifiers. *International Joint Conference on Neural Networks*, Killarney, Ireland, July 12-17, pp. 1-8, 2015.

[FDN01] Farrar, C.R.; Doebling, S.W.; Nix, D.A.: Vibration-based structural damage identification. In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 359, no. 1778, pp. 131-149, 2001.

[FL09] Franco, A.; Nanni, L.: Fusion of classifiers for illumination robust face recognition. In: *Expert Systems with Applications*, vol. 36, pp. 8946-8954, 2009.

[Fri05] Fritzen, C.P.: Vibration-based structural health monitoring - concepts and applications. In: *Key Engineering Materials*, vols. 293-294, pp. 3-20, 2005.

[FW07] Farrar, C.R.; Worden, K.: An introduction to structural health monitoring. In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 365, no. 1851, pp. 303-315, 2007.

[GA10] Gandossi, L.; Annis, C.: Probability of detection curves: Statistical best-practices. ENIQ report 41. Office for official Publications of the European Communities, Luxembourg, 2010.

[GR01] Giacinto, G.; Roli, F.: Design of effective neural network ensembles for image classification purposes. In: *Image and Vision Computing*, vol. 19, no. 9-10, pp. 699-707, 2001.

[Hal+09] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H.: The WEKA data mining software: An update. In: *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[HB11] Hou, L.; Bergmann, N.W.: Induction motor fault diagnosis using industrial wireless sensor networks and Dempster-Shafer classifier fusion. *37th Annual Conference on IEEE Industrial Electronics Society*, Melbourne, Australia, November 7-10, pp. 2992-2997, 2011.

[HHS94] Ho, T.K.; Hull, J.J.; Srihari, S.N.: Decision combination in multiple classifier systems. In: *IEEE Transactions on Pattern 495 Analysis and Machine Intelligence*, vol. 16, no. 1, pp. 66-75, 1994.

[HS93] Huang, Y.S.; Suen, C.Y.: The behavior-knowledge space method for combination of multiple classifiers. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, New York, USA, June 15-17, pp. 347-352, 1993.

[HVY11] Haghighi, M.S.; Vahedian, A.; Yazdi, H.S.: Extending Dempster Shafer method by multilayer decision template in classifier fusion. *7th International Conference on Information Assurance and Security*, Malacca, Malaysia, December 5-8, pp. 128-133, 2011.

[IB97] Isermann, R.; Balle, P.: Trends in the application of model-based fault detection and diagnosis of technical processes. In: *Control engineering practice*, vol. 5, no. 5, pp. 709-719, 1997.

[KBD01] Kuncheva, L.I.; Bezdek, J.C.; Duin, R.P.W.: Decision templates for multiple classifier fusion: an experimental comparison. In: *Pattern Recognition*, vol. 34, no. 2, pp. 99-314, 2001.

[KN05] Ko, J.M.; Ni, Y.Q.: Technology developments in structural health monitoring of large-scale bridges. In: *Engineering structures*, vol. 27, no. 12, pp. 1715-1725, 2005.

[KN14] Kampmann, G.; Nelles, O.: One-Class LS-SVM with Zero Leave-One-Out Error. In: *IEEE Symposium Series On Computational Intelligence*, Orlando, USA, December 9-12, pp. 1-6, 2014.

[Koh95] Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *IJCAI95 Proceedings of the 14th international joint conference on Artificial intelligence*, Montreal, Canada, August 20-25, vol. 2, pp. 1137-1145, 1995.

[KR14] Kuncheva, L.I.; Rodríguez, J.J.: A weighted voting framework for classifiers ensembles. In: *Knowledge and Information Systems*, vol. 38, no. 2, pp. 259-275, 2014.

[Krs+14] Krstajic, D.; Buturovic, L.J.; Leahy, D.E.; Thomas, S.: Cross-validation pitfalls when selecting and assessing regression and classification models. In: *Journal of Cheminformatics*, vol. 6, no. 1, pp. 10, 2014.

[KSB08] Ko, A.H.; Sabourin, R.; Britto Jr., A.S.: From dynamic classifier selection to dynamic ensemble selection. In: *Pattern Recognition*, vol. 41, no. 5, pp. 1718-1731, 2008.

[Kun04] Kuncheva, L.I.: *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, Inc., vol. 2, 2004.

[Kur+13] Kurz, J.H.; Jüngert, A.; Dugan, S.; Dobmann, G.; Boller,C.: Reliability considerations of NDT by probability of detection (POD) determination using ultrasound phased array. In: *Engineering Failure Analysis*, vol. 35, pp. 609-617, 2013.

[Li+13] Li, L.; Zou, B.; Hu, Q.; Wu, X.; Yu, D.: Dynamic classifier ensemble using classification confidence. In: *Neurocomputing*, vol. 99, pp. 581-591, 2013.

[Lin+14] Lin, C.; Chen, W.; Qiu, C.; Wu, Y.; Krishnan, S.; Zou, Q.: LibD3D: Ensemble classifiers with a clustering and dynamic selection strategy. *Neurocomputing*, vol. 123, pp. 424-435, 2014.

[Liu+18] Liu, Z.G.; Pan, Q.; Dezert, J.; Martin, A.: Combination of classifiers with optimal weight based on evidential reasoning. In: *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 3, pp. 1217-1230, 2018.

[Man+11] Mandache, C.; Genest, M.; Khan, M.; Mrad, N.: Considerations on structural health monitoring reliability. *International Workshop Smart Materials, Structures & NDT in Aerospace*, Montreal, Canada, November 2-4, 2011.

[Man$^+$15] MANALLAH, N.; ALKHALIFAH, A.; KHAN, R.; RAHMAN, H.U.; KHAN, S.: On the performance analysis of classifier fusion for land cover classification. *7th International Conference on Recent Advances in Space Technologies*, Istanbul, Turkey, June 16-19, pp. 271-275, 2015.

[MB13] MAJNIK, M.; BOSNIĆ, Z.: ROC analysis of classifiers in machine learning: A survey. In: *Intelligent Data Analysis*, vol. 17, no. 3, pp. 531-558, 2013.

[ME15] MOUSAVI, R.; EFTEKHARI, M.: A new ensemble learning methodology based on hybridization of classifier ensemble selection approaches. In: *Applied Soft Computing*, vol. 37, pp. 652-666, 2015.

[Moo$^+$15] MOOSAVIAN, A.; KHAZAEE, M.; NAJAFI, G.; KETTNER, M.; MAMAT, R.: Spark plug fault recognition based on sensor fusion and classifier combination using Dempster-Shafer evidence theory. In: *Applied Acoustics*, vol. 93, pp. 120-129, 2015.

[Mor$^+$06] MORENO-SECO, F.; IESTA, J.M.; DE LEN, P.J.P.; MIC, L.: Comparison of classifier fusion methods for classification in pattern recognition tasks. In: Yeung, D. Y.; Kwok, J. T.; Fred, A.; Roli, F.; de Ridder D. (Eds.), *Structural, Syntactic, and Statistical Pattern Recognition. SSPR /SPR 2006. Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, pp. 705-713, 2006.

[MST94] MICHIE, D.; SPIEGELHALTER, D.J.; TAYLOR, C.C.: *Machine learning, neural and statistical classification*, Ellis Horwood Ltd, Publisher, 1994.

[MYL13] MA, A.J.; YUEN, P.C.; LAI, J.: Linear dependency modeling for classifier fusion and feature combination. In: *IEEE Transactions on pattern analysis and Machine Intelligence*, vol. 35, pp. 1135-1148, 2013.

[MZB12] MIKHAIL, M.; ZEIN-SABATTO, S.; BODRUZZAMAN, M.: Decision fusion methodologies in Structural Health Monitoring systems. In: *Proceedings of IEEE Southeastcon*, Orlando, USA, March 15-18, pp. 1-6, 2012.

[Ngu$^+$14] NGUYEN, T.T.; LIEW, A.W.C.; PHAM, C.X.; NGUYEN, M.P.: Optimization of ensemble classifier system based on multiple objectives genetic algorithm. In: *Proceedings of International Conference on Machine Learning and Cybernetics*, Lanzhou, China, July 13-16, pp. 46-51, 2014.

[Ngu^{+}18] Nguyen, T.T.; Pham, C.X.; Liew, A.W.C.; Pedrycz, W.: Aggregation of classifiers: A justifiable information granularity approach. In: *IEEE Transactions on Cybernetics*, vol. 99, pp. 1-10, 2018.

[NL09] Nanni, L.; Lumini, A.: A genetic encoding approach for learning methods for combining classifiers. In: *Expert Systems with Applications*, vol. 36, no. 4, pp. 7510-7514, 2009.

[NN12] Nabiha, A.; Nadir, F.: New Dynamic Ensemble of Classifiers Selection approach based on confusion matrix for Arabic Handwritten recognition. *International Conference on Multimedia Computing and Systems*, Tangier, Morocco, May 10-12, pp. 308-313, 2012.

[Ouk^{+}10] Oukhellou, L.; Debiolles, A.; Denæux, T.; Aknin, P.: Fault diagnosis in railway track circuits using Dempster-Shafer classifier fusion. In: *Engineering Applications of Artificial Intelligence*, vol. 23, pp. 117-128, 2010.

[PG11] Paksoy, A.; Göktürk, M.: Information fusion with dempster-shafer evidence theory for software defect prediction. In: *Procedia Computer Science*, vol. 3, pp. 600-605, 2011.

[PGO14] Pérez, M.A.; Gil, L.; Oller, S.: Impact damage identification in composite laminates using vibration testing. In: *Composite Structures*, vol. 108, no. 1, pp. 267-276, 2014.

[Poo^{+}17] Poon, J.; Jain, P.; Konstantakopoulos I.C.; Spanos C.; Panda S.K.; Sanders S.R.: Model-based fault detection and identification for switching power converters. In: *IEEE Transactions on Power Electronics*, vol. 32, no. 2, pp. 1419-1430, 2017.

[QMD11] Quost, B.; Masson, M.-H.; Denæux, T.: Classifier fusion in the Dempster-Shafer framework using optimized t-norm based combination rules. In: *International Journal of Approximate Reasoning*, vol. 52, pp. 353-374, 2011.

[QZG16] Qu, J.; Zhang, Z.; Gong, T.: A novel intelligent method for mechanical fault diagnosis based on dual-tree complex wavelet packet transform and multiple classifier fusion. In: *Neurocomputing*, vol. 171, pp. 837-853, 2016.

[RG00] Ruta, D.; B. Gabrys: An overview of classifier fusion methods. In: *Computing and Information Systems*, vol. 7, pp. 1-10, 2000.

[RJS14] Rother, A.; Jelali, M.; Söffker, D.: Development of a fault detection approach based on SVM applied to industrial data. In: *Proceedings of EWSHM -7th European Workshop on Structural Health Monitoring*, Nantes, France, July 8-11, 2014.

[RJS15] Rother, A.; Jelali, M.; Söffker, D.: A brief review and a first application of time-frequency-based analysis methods with application to strip rolling mills. In: *Journal of Process Control*, vol. 35, pp. 65-79, 2015.

[Rot$^+$17] Rothe, S.; Wirtz, S.F.; Kampmann, G.; Nelles, O.; Söffker, D.: Ensure the reliability of damage detection in composites by fusion of differently classified Acoustic Emission measurements. In: Chang, F.K.; Kopsaftopoulos, F. (Ed.): *Structural Health Monitoring 2017*, Stanford, USA, September 12-14, 2017, pp. 1380-1387.

[RPL10] Rodriguez, J.D.; Perez, A.; Lozano, J. A.: Sensitivity analysis of k-fold cross validation in prediction error estimation, In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 569-575, 2010.

[RS16] Rothe, S.; Söffker, D.: Comparison of different information fusion methods using ensemble selection considering benchmark data. In *Proceedings of 19th International Conference on Information Fusion (FUSION)*, Heidelberg, Germany, July 5-8, pp. 73-78, 2016.

[RS19] Rothe, S.; Söffker, D.: Does the precision value influence the fusion performance? A method-based experimental study. Reliability Engineering & System Safety, 2019, submitted.

[RKS19] Rothe, S.; Kudszus, B.; Söffker, D.: Does classifier fusion improve the overall performance? Numerical analysis of data and fusion method characteristics influencing classifier fusion performance. Information Fusion, 2019, submitted.

[RWS16] Rothe, S.; Wirtz, S. F.; Söffker, D.: About the reliability of diagnostic statements: fundamentals about detection rates, false alarms, and technical requirements. *11. Aachener Kolloquium für Instandhaltung, Diagnose und Anlagenüberwachung*, Aachen, Germany, November 15-16, 2016.

[Sch$^+$04] Schenker, A.; Bunke, H.; Last, M.; Kandel, A.: Building graph-based classifier ensembles by random node selection. In: Roli F., Kittler J., Windeatt T. (eds) *Multiple Classifier Systems. MCS 2004. Lecture Notes in Computer Science*, vol. 3077, pp. 214-222, 2004.

[Sch+15] Schubert Kabban, C.M.; Greenwell, B.M.; DeSimio, M.P.; Derriso, M.M.: The probability of detection for structural health monitoring systems: Repeated measures data. In: *Structural Health Monitoring*, vol. 14, no. 3, pp. 252-264, 2015.

[Sha76] Shafer, G.: *A mathematical theory of evidence.* Princeton: Princeton University Press, 1976.

[Sha+00] Sharkey, A.J.C.; Sharkey, N.E.; Gerecke, U.; Chandroth, G.O.: The "test and select" approach to ensemble combination. In: *Multiple Classifier Systems 2000*, Cagliari, Italy, June 2123, 2000.

[SL87] Shafer, G.; Logan, R.: Implementing dempsters rule for hierarchical evidence. In: *Artificial Intelligence*, vol. 33, no. 3, pp. 271-298, 1987.

[SL00] Suen, C.Y.; Lam, L.: Multiple classifier combination methodologies for different output levels, In: *Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, vol. 1857, pp. 52-66, 2000.

[Soa+06] Soares, R.G.F.; Santana, A.; Canuto, A.M.P.; de Souto, M.C.P.: Using accuracy and diversity to select classifiers to build ensembles. *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, Vancouver, Canada, pp. 1310-1316, 2006.

[Soe+16] Söffker, D.; Wei, C.; Wolff, S.; Saadawia, M.S.: Detection of rotor cracks: comparison of an old model-based approach with a new signal-based approach. In: *Nonlinear Dynamics*, vol. 83, no. 3, pp. 1153-1170, 2016.

[Son14] Song, L: A NSGA-II program in Matlab v1.4. http://www.mathworks.com/matlabcentral/fileexchange/31166-ngpm-a-nsga-ii-program-in-matlab-v1-4. Access date: 20.02.2014.

[Sue90] Suen, C.Y.: Recognition of totally unconstrained handwritten numerals based on the concept of multiple experts. In: *Proc. International Workshop on Frontiers in Handwriting Recognition*, Montreal, Canada, April 23, pp. 131-143, 1990.

[Tam+11] Tamminedi, T.; Ganapathy, P.; Zhang, L.; Yadegar, J.: Classifier fusion framework using genetic algorithms. In: *IEEE 22nd International Symposium on Personal Indoor and Mobile Radio Communications*, Toronto, Canada, September 11-14, pp. 2224-2228, 2011.

[Tul+08] Tulyakov, S.; Jaeger, S.; Govindaraju, V.; Doermann, D.: Review of classifier combination methods. In: Marinai, S.; Fujisawa, H. (Eds.) *Machine Learning in Document Analysis and Recognition*, Springer Berlin Heidelberg, vol. 90, pp. 361-386, 2008.

[Vri+15] Vriesmann, L.M.; Britto Jr, A.S.; Oliveira, L.S.; Koerich, A.L.; Sabourin, R.: Combining overall and local class accuracies in an oracle-based method for dynamic ensemble selection. *International Joint Conference on Neural Networks*, Killarney, Ireland, July 12-17, pp. 1-7, 2015.

[VS06] Varma, S.; Simon, R.: Bias in error estimation when using cross-validation for model selection. In: *BMC Bioinformatics*, vol. 7, no. 91, pp. 1-8, 2006.

[Wan+98] Wang, D.; Keller, J. M.; Carson, C.A.; McAdoo-Edwards, K.K.; Bailey, C.W.: Use of fuzzy-logic-inspired features to improve bacterial recognition through classifier fusion. In: *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 28B, no. 4, pp. 583-591, 1998.

[Wan+14] Wang, X.; Wang, R.; Feng, H.; Wang, H.: A new approach to classifier fusion based on upper integral. In: *IEEE Transactions on Cybernetics*, vol. 44, pp. 620-635, 2014.

[WB06] Walt, C.V.D.; Barnard, E.: Data characteristics that determine classifier performance. In: *SAIEE Africa Research Journal*, vol. 98, no. 3, pp. 87-93, 2006.

[WBS18] Wirtz, S.; Beganovic, N.; Söffker, D.: Investigation of damage detectability in composites using frequency-based classification of Acoustic Emission measurements. In: *Structural Health Monitoring*, pp. 1-12, 2018.

[WM97] Wolpert, D.H.; Macready, W.G.: No free lunch theorems for optimization. In: *IEEE Transactions on Evolutionary Computation*, vol. 1, pp. 67-82, 1997.

[Won15] Wong, T.T.: Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation In: *Pattern Recognition*, vol. 48, no. 9, pp. 2839-2846, 2015.

[WTG09] Wu, Y.; Tan, X.; Gu, S.: A learning evaluation system based on classifier fusion for E-learning. *IEEE International Symposium on IT in Medicine & Education*, Jinan, China, August 14-16, vol. 1, pp. 749-752, 2009.

[WY07] Widodo, A.; Yang, B.S.: Support vector machine in machine condition monitoring and fault diagnosis. In: *Mechanical systems and signal processing*, vol. 21, no. 6, pp. 2560-2574, 2007.

[XKH11] Xia, M.; Kong, F.; Hu, F.: An approach for bearing fault diagnosis based on PCA and multiple classifier fusion. *6th IEEE Joint International Information Technology and Artificial Intelligence Conference*, Chongqing, China, August 20-22, vol. 1, pp. 321-325, 2011.

[XKS92] Xu, L.; Krzyzak, A.; Suen, C.Y.: Methods of combining multiple classifiers and their applications to handwriting recognition. In: *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, no. 3, pp. 418-435, 1992.

[Yan11] Yang, L.: Classifiers selection for ensemble learning based on accuracy and diversity. In: *Procedia Engineering*, vol. 15, pp. 4266-4270, 2011.

[ZD11] Zhang, C.; Duin, R.P.W.: An experimental study of one- and two-level classifier fusion for different sample sizes. Pattern Recognition Letters, vol. 32, pp. 1756-1767, 2011.

[Zhe15] Zheng, A.: *Evaluating machine learning models*, OReilly Media, Inc., USA, 2015.

This thesis is based on the results and development steps presented in the following previous publications:

Journal articles

[RS19] Rothe, S.; Söffker, D.: Does the precision value influence the fusion performance? A method-based experimental study. Reliability Engineering & System Safety, 2019, submitted.

[RKS19] Rothe, S.; Kudszus, B.; Söffker, D.: Does classifier fusion improve the overall performance? Numerical analysis of data and fusion method characteristics influencing classifier fusion performance. Information Fusion, 2019, submitted.

[ARS19] Ameyaw, D. A.; Rothe, S.; Söffker, D.: A novel feature-based Probability of Detection (POD) assessment and fusion approach for reliability evaluation of vibration-based SHM systems. SHM Journal, 2019, accepted.

Conference papers

[ARS18a] Ameyaw, D. A.; Rothe, S.; Söffker, D.: Adaptation and implementation of Probability of Detection (POD)-based fault diagnosis in elastic structures through vibration-based SHM approach. The 9th European Workshop on Structural Health Monitoring (EWSHM) , Manchester, July 10-13, 2018.

[ARS18b] Ameyaw, D. A.; Rothe, S.; Söffker, D.: Probability of Detection (POD)-oriented view to fault diagnosis for reliability assessment of FDI approaches. ASME 2018 International Design Engineering Technical Conferences & Computers. 30th Conference on Mechanical Vibration and Noise, Quebec City, August 26-29, 2018, DETC2018-85554, pp.V008T10A041.

[ARS18c] Ameyaw, D. A.; Rothe, S.; Söffker, D.: Fault diagnosis using Probability of Detection (POD)-based sensor/information fusion for vibration-based analysis of elastic structures. PAMM-Wiley Online, 2018.

[Rot$^+$17] Rothe, S.; Wirtz, S.F.; Kampmann, G.; Nelles, O.; Söffker, D.: Ensure the reliability of damage detection in composites by fusion of differently classified Acoustic Emission measurements. In: Chang, F.K.; Kopsaftopoulos, F. (Eds.): Structural Health Monitoring 2017, Stanford, USA, September 12-14, pp. 1380-1387, 2017.

[RS16] Rothe, S.; Söffker, D.: Comparison of different information fusion methods using ensemble selection considering benchmark data. 19th International Conference on Information Fusion (FUSION), Heidelberg, Germany, July 5-8, pp. 73-78, 2016.

[RWS16] Rothe, S.; Wirtz, S. F.; Söffker, D.: About the reliability of diagnostic statements: fundamentals about detection rates, false alarms, and technical requirements. 11. Aachener Kolloquium für Instandhaltung, Diagnose und Anlagenüberwachung, Aachen, Germany, November 15-16, 2016.

Other publications, which are not included in this thesis:

Journal articles

[SR17] Söffker, D.; Rothe, S.: New Approaches for Supervision of Systems with Sliding Wear: Fundamental Problems and Experimental Results Using Different Approaches. Applied Sciences, vol. 7, pp. 843, 2017.

Conference papers

[BRS16] Beganovic, N.; Rothe, S.; Söffker, D.: Identification of diagnostic and prognostic features by means of AE and hydraulic pressure measurements. European Workshop on Structural Health Monitoring, Bilbao, Spain, July 5-8, 2016.

[RS15] Rothe, S.; Söffker, D.: Development of a state-related evaluation for diagnostic-oriented data filtering approach. In: Chang, F.K.; Kopsaftopoulos, F. (Eds.): Structural Health Monitoring 2015, Stanford, USA, September 1-3, pp. 593-600, 2015.

[Rot$^+$15] Rothe, S.; Leite, A.; Padrao, P.; Söffker, D.: Improvement and comparison of wear-oriented state-of- health classification methods using optimization techniques. In: Chang, F.K.; Kopsaftopoulos, F. (Eds.): Structural Health Monitoring 2015, Stanford, USA, September 1-3, pp. 625-632, 2015.

[Beg$^+$15] Beganovic, N.; Njiri, J.G.; Rothe, S.; Söffker, D.: Application of diagnosis and prognosis to wind turbine system based on fatigue load. Proc. IEEE Conference on Prognostics and Health Management PHM 2015, Austin, TX, USA, pp. 1-6, 2015.

[Pad$^+$15] Padrao, P.; Rothe, S.; Leite, A.; Söffker, D.: Optimal threshold synthesis for State-of-Health classification and evaluation of a tribological system. Proc. 17th International Symposium on Dynamic Problems of Mechanics, Natal, Brazil, February 22-27, 2015.

[RS14] Rothe, S.; Söffker, D.: Wear-oriented state-of-health calculation and classification using operating data. Proc. Le Cam, Vincent and Mevel, Laurent and Schoefs, Franck. EWSHM -7th European Workshop on Structural Health Monitoring, Nantes, France, July 8-11, 2014.

[SR14] Söffker, D.; Rothe, S.: Comparison of three - easy to apply and innovative - signal-based approaches for diagnosis of a technical system with wear. Proc. ASME 2014 Dynamic Systems and Control (DSC) Conference, vol. 1, pp. V001T08A005, 2014.

[BRS14] Beganovic, N.; Rothe, S.; Söffker, D.: Establishing a wear-related deterioration model based on experimental data. Proc. Le Cam, Vincent and Mevel, Laurent and Schoefs, Franck. EWSHM -7th European Workshop on Structural Health Monitoring, Nantes, France, July 8-11, 2014.

[SBR14] Söffker, D.; Beganovic, N.; Rothe, S.: Von der Diagnose zur Prognose: signal- und modellbasierte Methoden zur aktiven Anlagenüberwachung am Beispiel der berwachung eines Verschleißprozesses. 10. Aachener Kolloquium für Instandhaltung, Diagnose und Anlagenüberwachung (AKIDA), Aachen, pp. 293-303, 2014.

[Sch$^+$14] Schiffer, S.; Rothe, S.; Baccar, D.; Söffker, D.: Classifiation of system's health condition using the new Adaptive Fuzzy-based Feature Classification Approach AFFCA in comparison to a macro-data-based approach. Proc. Le Cam, Vincent and Mevel, Laurent and Schoefs, Franck. EWSHM -7th European Workshop on Structural Health Monitoring, Nantes, France, July 8-11, 2014.

[Soe13] Söffker, D.; Rothe, S.; Schiffer, S.; Aljoumaa, H.; Baccar, D.: Smart, tough, and successful: Three new innovative approaches for diagnosis and prognosis of technical systems. In: Chang, F.K. (Ed.): Structural Health Monitoring 2013, pp. 81-88, 2013.

[Soe12] Söffker, D.; Aljoumaa, H.; Baccar, D.; Rothe, S.: Smart, robust und einfach: Drei innovative Konzepte zur Maschinendiagnose. 9. Aachener Kolloquium für Instandhaltung, Diagnose und Anlagenüberwachung AKIDA, Aachen, 2012.

[Soe$^+$14] Söffker, D.; Muthig O.; Hägele, G.; Sarkheyli, A.; Rothe, S.: Automat oder/und Mensch - Assistenz, Führung, wechselnde Rollenverteilung: berlegungen zur Berechnung und Steigerung der Gesamtsystemverlässlichkeit. Tagungsband DGLR L6.4 Anthropotechnik 56. Fachausschusssitzung, Ottobrunn, Deutschland, Oktober 14-15, pp. 281-284, 2014.

In the context of research projects at the Chair of Dynamics and Control, the following student thesis has been supervised by Sandra Rothe and Univ.-Prof. Dr.-Ing. Dirk Söffker. Development steps and results of the research work and the student theses are integrated with each other and hence are also part of this thesis:

[Kud18] Kudszus, B., Analysis of relationships between different data and classifier properties and related usability of fusion methods, Master Thesis, September 2018.

The following student theses have been supervised by Sandra Rothe and Univ.-Prof. Dr.-Ing. Dirk Söffker, which are not included in this thesis:

[Yil18] Yildiz, T.N., Diagnose und Prognose von Maschinendaten mit Hilfe von Big Data Analytics, Bachelor Thesis, March 2018.

[Mey16] Meyer, D., Inbetriebnahme und Optimierung eines neuen Verschleißversuchsstandes zur Bestimmung des Funktionsverlustes von Reibkontakten, Bachelor Thesis, February 2016.

[Lie15] Lietz, M., Analyse, Umkonstruktion und Umbau eines bestehenden Prüfstandes für Verschleißversuche, Master Thesis, August 2015.

[Li14] Li, C., Literature research and application of evaluation systems for condition monitoring, Bachelor Thesis, August 2014.

DOI: 10.17185/duepublico/70209
URN: urn:nbn:de:hbz:464-20190624-101115-2