New parametric evaluation and fusion strategy for vibration diagnosis systems and classification approaches applied to machine learning and computer vision systems

Von der Fakultät für Ingenieurwissenschaften,
Abteilung Maschinenbau und Verfahrenstechnik der
Universität Duisburg-Essen
zur Erlangung des akademischen Grades
eines
Doktors der Ingenieurwissenschaften
Dr.-Ing.
genehmigte Dissertation

von

Daniel Adofo Ameyaw
aus
Ghana

Gutachter:
Univ.-Prof. Dr.-Ing. Dirk Söffker
Univ.-Prof. Dr.-Ing. Peter Kraemer

Tag der mündlichen Prüfung: 30. Juli 2020

Acknowledgements

This work was accomplished with the support of research collaboration between the Ministry of Education, Ghana and the German Academic Exchange Service (DAAD).

I would like to express my appreciation to my supervisor Univ.-Prof. Dr.-Ing. Dirk Söffker, for granting me the opportunity to accomplish my research at the Chair of Dynamics and Control (SRS) at University of Duisburg-Essen. I really appreciate his guidance, counsel, and help.

I would like to express my gratitude to Univ.-Prof. Dr.-Ing. Peter Kraemer for his effort and time in examining my thesis.

My appreciation to colleagues at the Chair of Dynamics and Control for their support.

Heartfelt gratitude to my family and loved ones for their support. Your thoughtfulness is a gift I will always treasure.

Duisburg, August 2020 Daniel Adofo Ameyaw

Kurzfassung

Fortschritte in Wissenschaft und Technik haben zur Entwicklung komplexer, sicherheitskritischer und kapitalintensiver Systeme und Prozesse geführt. Die Gewährleistung eines optimalen und sicheren Betriebs dieser technischen Systeme und Prozesse erfordert die Entwicklung von Konzepten zur Qualitätsüberwachung und Kontrolle. Die während der Überwachung und Aufsicht gewonnenen Daten können zur Diagnose des Systemzustands ausgewertet werden. Die Auswahl eines Evaluierungsansatzes für eine Anwendung kann eine Leistungsbewertung erfordern. In dieser Arbeit wird ein neuer Evaluierungsansatz entwickelt, der auf dem Zuverlässigkeitsmaß Probability of Detection (POD) basiert. Die POD ist eine probabilistische Methode zur Quantifizierung der Zuverlässigkeit eines Verfahrens unter Berücksichtigung der statistischen Variabilität der Sensor- und Messeigenschaften. Das $a_{90/95}$-Kriterium, das die Wahrscheinlichkeit einer Erkennung von 90 % bei einem Konfidenzniveau von 95 % darstellt, wird mit Messungen und den damit verbundenen Ergebnissen untersucht. Die Verallgemeinerungsfähigkeit des entwickelten Ansatzes wird in multidisziplinären Bereichen demonstriert: schwingungsbasiertes Structural Health Monitoring (SHM), Maschinelles Lernen (ML) und Computer Vision (CV) Klassifizierungsansätze.

Structural Health Monitoring-Systeme basieren auf geeigneten Sensortechniken, die eine Online- und Offline-Überwachung von technischen Systemen ermöglichen. Für schwingungsbasierte Überwachungsansätze kann der POD-Ansatz nicht in ähnlicher Weise angewendet werden. Dies resultiert hauptsächlich aus der Komplexität des dynamischen Verhaltens der überwachten Systeme in Bezug auf Fehler, Sensorposition (Beobachtbarkeit) und der damit verbundenen Merkmalsextraktion oder Überwachungsaufgabe. In dieser Forschungsarbeit wird die POD-Bewertung der schwingungsbasierten Fehlererkennung von zu überwachenden elastisch mechanischen Strukturen entwickelt. Neben einer grundsätzlichen Diskussion des Problems, die als Einführung dient, wird ein Beispiel mit verschiedenen Sensortypen in Kombination mit mechanischen Modifikationen eines elastischen Balkens vorgestellt. Basierend auf der Analyse eines geeignet gewählten Merkmals und in Abhängigkeit von den betrachteten mechanischen Moden wird die Effizienz und der Mangel der verschiedenen Kombinationen aufgezeigt. Auf der Grundlage des vorgeschlagenen Ansatzes wird ein neuer Einblick in die Nützlichkeit verschiedener Kombinationen von Sensortyp und Fehlerposition möglich. Zur Verbesserung der Detektionsqualität werden geeignete Annahmen in Kombination mit der Sensor/Informationsfusion auf die merkmalsbasierte Analyse als Detektionsaufgabe mittels Schwingungsmessungen angewendet. Mit Hilfe der Rauschanalyse kann ein Kompromiss zwischen der Fehlalarmrate (PFP) und der mit einem Zuverlässigkeitswert von $a_{90/95}$ zu detektierenden Fehlergröe erreicht werden.

Ansätze des Maschinellen Lernens (ML) bieten die Möglichkeit, automatisch aus Erfahrungen zu lernen und sich zu verbessern. Die ML-Ansätze sind funktionale

Werkzeuge, die Klassifizierungsfähigkeiten ermöglichen. In diesem Sinn ist auch die auf Maschinellem Lernen basierende Erkennung komplexer Fahrsituationen sowohl für automatisierte Fahrzeuge als auch für menschliche Fahrer von zunehmendem Interesse. In diesem Bereich müssen auf der Basis technischer Sensoren ein komplexes Szenario und die damit verbundenen dynamischen Veränderungen interpretiert und klassifiziert werden. Die Bewertung dieser Klassifikatoren spielt eine entscheidende Rolle bei der Auswahl für eine bestimmte Aufgabe. In dieser Arbeit wird ein neuer Ansatz entwickelt, der einen neuen Bewertungsmastab für ML-Ansätze, die als Klassifikatoren verwendet werden, ermöglicht. Gegenwärtige Methoden verwenden Maße wie die Receiver Operating Characteristic (ROC)-Kurve, die visuell und grafisch die Leistung von Klassifikatoren darstellt. Neben dem Verhältnis von Erkennungsrate und Fehlalarmrate (kombiniert als ROC-Kurve) werden andere Eigenschaften, die sich auf Prozessparameter beziehen, nicht in den Evaluierungsprozess integriert. In dieser Forschung wird eine neue Evaluierungsmethode entwickelt, die auf dem Zuverlässigkeitsmaß Probability of Detection (POD) basiert und den Einfluss weiterer Prozessparameter auf die Klassifikationsergebnisse integriert und diskutiert. Als Beispiel zur Veranschaulichung wird der Vergleich von Klassifikatoren, die auf das Vorhersageverhalten von Fahrern angewendet werden, herangezogen. Der zeitliche Abstand der Entscheidungsmomente zum Zeitpunkt der Entscheidung selbst wird als Prozessparameter für den Klassifizierungsprozess verwendet. Der vorgeschlagene Ansatz erlaubt den Vergleich der Eignung von Künstlichen Neuronalen Netzen, Hidden-Markov-Modellen, Random Forest, Support Vector Machines und verbesserten Versionen dieser Klassifikatoren im Hinblick auf die Zuverlässigkeit der zugehörigen Klassifikation bevorstehender Ereignisse vor den realen Ereignissen, hier: menschliche Entscheidungen auf der Grundlage des visuellen Eindrucks von Veränderungen innerhalb der Umgebung. Folglich können auf der Grundlage der POD-bezogenen Auswertung die verschiedenen Klassifikatoren hinsichtlich ihrer Fähigkeit, das richtige Verhalten als Funktion der Zeit vor dem Ereignis selbst vorherzusagen, klar unterschieden werden.

Bei Computer Vision-Systemen ist die Bildklassifizierung und erkennung eine wichtige Aufgabe. Deep Convolutional Neural Network (CNN) ist einer der vielversprechenden Klassifikatoren fr die groß flächige Bildklassifikation. Es ist bekannt, dass die Leistung der Bildklassifikation stark von den Problemeigenschaften und den Prozessparametern abhängt. Diese Arbeit konzentriert sich auf die explizite und genaue Bewertung von CNN-Klassifikatoren bei der Bilderkennung. Lösungsvorschlag ist die Implementierung des POD-Ansatzes, der die Auswertung weiterer Bildparameter erlaubt, die die Klassifikationsergebnisse beeinflussen, und damit auch den Vergleich von Klassifikatoren, die sich je nach Klassifikationsaufgabe auf den gleichen Parameter beziehen. Die vorgeschlagene Evaluierungsmethode wird an einer Reihe von Klassifikatoren implementiert und Vergleiche von Parametern mit den besten Detektionsfähigkeiten werden durchgeführt.

Abstract

Advancements in science and technology has led to development of complex, safety-critical, and capital-intensive systems and processes. To ensure optimal and safe operations of these technical systems and processes requires the development of quality monitoring and supervision approaches. Data obtained during monitoring and supervision can be evaluated for state-of-health diagnosis. The selection of an evaluation approach for an application may require performance assessment. In this thesis a new evaluation approach based on the Probability of Detection (POD) reliability measure is developed. The POD is a probabilistic method to quantify the reliability of a procedure taking into account statistical variability of sensor and measurements properties. The $a_{90/95}$-criteria representing probability of 90 % detection at a confidence level of 95 % is examined to the measurements and related outcomes. The generalizable capability of the developed approach is demonstrated in multidisciplinary fields: vibration-based Structural Health Monitoring (SHM), Machine Learning (ML), and Computer Vision (CV) classification approaches.

Structural Health Monitoring systems are based on suitable sensor techniques allowing online and offline supervision of technical systems. The quantification of sensors/measurement devices is a key issue for qualifying their effectiveness and efficiency. For vibration-based supervision approaches the POD approach can not be applied similarly. This results mainly from the complexity of the dynamical behavior of systems monitored in relation to faults, sensors position (observability), and the related feature extraction or monitoring task. In this research, POD evaluation of vibration-based fault detection of elastic mechanical structures to be monitored is developed. Beside a principal discussion of the problem serving as introduction, an example using different sensor types in combination with mechanical modifications of an elastic beam is presented. Based on the analysis of a suitably chosen feature and dependent on the mechanical modes considered, the efficiency and deficiency of the different combinations are shown. Based on the proposed approach, a new insight to the usefulness of different sensor types and fault position combination becomes possible. To improve the detection quality, suitable assumptions in combination with sensor/information fusion are applied to feature-based analysis as detection task using vibration measurements. Dependent on noise analysis, a trade-off between flaw size detection and probability of falsely characterizing a fault with $a_{90/95}$ reliability level can be attained.

Machine learning approaches provide the ability to automatically learn and improve from experiences. These ML approaches are functional tools that allow classification abilities. In this sense also the machine learning-based recognition of complex driving situations for automated vehicles as well as for human drivers is of increasing interest. In this field based on technical sensors a complex scenario and related dynamical changes have to be interpreted and classified. Evaluation of these classifiers plays a crucial role in their selection for a specific task. In this thesis a new approach

is established permitting a new evaluation measure to ML approaches used as classifiers. Current methods utilize measures like the receiver operating characteristic (ROC) providing visually and graphically the performance of classifiers. Beside the ratio of detection rate and false alarm rate (combined as ROC), other properties related to process parameters are not integrated in the evaluation process. In this research, a new evaluation method based on the Probability of detection (POD) reliability measure is developed integrating and discussing the effect of further process parameters on the classification results. As an example for illustration the comparison of classifiers applied to driver prediction behavior is used. The temporal distance of decisions moments to the instant of decision itself is used as process parameter for the classification process. The proposed approach is used to compare the suitability of Artificial Neural Networks, Hidden Markov Models, Random Forest, Support Vector Machines, and improved versions of these classifiers with respect to the reliability of related classification of upcoming events prior to the real ones, here: human decisions based on the visual impression of changes within the environment. Consequently, based on the POD-related evaluation the different classifiers can be clearly distinguished with respect to their ability to predict the correct behavior as a function of time prior to the event itself.

In Computer Vision systems, image classification and detection is an important task. Deep Convolutional Neural Network (CNN) is a suitable classifier for large-scale image classification. It is well-known that image classification performance strongly depends on problem characteristics and vary with process parameters. This work focuses on explicit and accurate evaluation of CNN classifiers in image detection. The POD approach is implemented allowing the evaluation of further image parameters affecting the classification results and subsequently the comparison of classifiers related to the same parameter depending on classification task. The proposed evaluation method is implemented on a number of classifiers and comparisons made on parameters with the best detection capabilities.

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
BKS_l	Unit l of the Behavior Knowledge Space
β_0	Ordinate intercept
β_1	Gradient
$\hat{\beta}_0, \hat{\beta}_1$	Maximum Likelihood estimation for β_0, β_1
χ^2	Chi squared hypothesis test
C_i	Class i with $i = 1, ..., n_c$
C^k	Confusion matrix for classifier e_k
C^k_{ij}	Number of samples, where classifier e_k assigns class j and actual class i
$E(x)$	Expected value
erf	Error function
$f(X)$	Algebraic function with linearized parameters
$g(y)$	Link function
$\hat{\mu}_{noise}$	Mean of noise data
$\hat{q}$	Fused class label
$\hat{\sigma}_{noise}$	Standard deviation of noise data
I	Information matrix
μ	Mean
n_{hits}	Number of hits
n_k	Number of classifiers
n_{miss}	Number of targets missed
$n_{targets}$	Number of observations considered in the POD evaluation process
n_{total}	Total number of observations
$P1, P2, P3$	Sensors at position 1, 2, 3
P	Probability
Φ	Cumulative distribution function
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
P_1, P_2, P_3	Probabilities of Sensor 1, 2, 3
q_k	Assigned class by classifier e_k

S_1, S_2, S_3	Lane change behavior
σ	Standard deviation for POD curve
τ	Standard deviation for regression line
$y_{a=0.95}$	95 % Wald confidence bounds on y
$Y(i)$	Response variable
y_{th}	Decision threshold
z	Standardized deviation of regression line

Abbreviations

ACC	Accelerometer sensor
AIC	Akaike Information Criterion
ANN	Artificial Neural Network
AP	Average Precision
AUC	Area Under ROC Curve
AUPRC	Area Under PR Curve
BCC	Beyesian Combination Rule
BKS	Behavior Knowledge Space
CDF	Cumulative Density Function
CNN	Convolutional Neural Network
CFRP	Carbon Fiber Reinforced Polymer
DET	Detection-Error Tradeoff Curve
DR	Detection Rate
FAR	False Alarm Rate
FDI	Fault Detection and Isolation
FM	F-Measure/Score
FN	False Negative
FP	False Positive
FRR	False Rejection/Recognition Rate
FT	Fourier Transform
GLM	Generalized Linear Model
HMM	Hidden Markov Model
Laser	Laser sensor
mAP	Mean Average Precision
MAPOD	Model Assisted Probability of Detection
MLE	Maximum Likelihood estimation
MV	Majority Voting
NDT	Nondestructive Testing
NPV	Negative Predictive Value
PFP	Probability of False Positive
PM	Performance Measure
POD	Probability of Detection

PPV	Positive Predictive Value
PR	Precision Recall
RF	Random Forest
RMS	Root Mean Square
ROC	Receiver Operating Characteristic
SG	Strain Gauge sensor
SHM	Structural Health Monitoring
STFT	Short Time Fourier Transform
SVM	Support Vector Machine
TN	True Negative
TP	True Positive

1 Introduction

1.1 Motivation and problem statement

Faults, defects, and local changes present in components and systems may affect its structural and functional integrity. From this, it can be concluded that the structural/functional integrity and the quality of detection and diagnosis approaches are connected. Ensuring structural and functional safety guarantee optimal functioning. The aim of Fault Detection and Isolation (FDI) techniques and Structural Health Monitoring (SHM) approaches are to ensure optimal system performance. Fault Detection and Isolation methods can primarily be grouped into four categories: signal-based, model-based, data driven, and hybrid approaches. Structural health monitoring on the other hand is the process of implementing a damage detection and characterization strategy for monitoring engineering structures [KN05]. Several FDI and SHM procedures exist with vibration-based health monitoring a well-known and established approach. Vibration-based diagnosis use structural behavior for fault diagnosis and its based on the principle that structural defects result in changes in dynamical properties. Fault detection is usually based on displacement, velocity, or acceleration measurements at a point on the component/system [Fri05]. Although useful analysis may consider multipoint measurements and observable effects [ARS20].

Monitoring and supervision of vibrating structures is complex. The complexity is related to data acquisition and cleansing, feature extraction, and/or statistical modeling for feature classification [FDN01] [FDDN99]. Data acquisition is mainly based on hardware and software systems and defines the data to be acquired, the transmission medium, and the storage devices [SFHW01] [FDDN99]. Data cleansing involves data normalization procedures. Sensors form an integral part of the hardware system. Selection of appropriate types of sensors, quantities, and locations need to be effectively addressed. This ensures optimal sensor placement and prevent redundancy. Feature extraction details the selection and processing of features sensitive to a specific fault type. This is important due to different sensitivity levels of features to different types of faults. A suitable solution may be to select different features and conclude on the most sensitive to a fault present. Feature selection also addresses statistical distribution of features and data condensation [FDDN99]. The development of statistical models for change/fault classification completes the process in many instances though other processes may be implemented. Monitoring dynamic systems requires all outlined phases to be examined in a consistent and coherent manner [ARS20], however techniques to accomplish this needs to be developed.

Decision making for diagnosis systems requires the collection of data from the system to be monitored and proper data interpretation. This requires data to be processed

to easily understood form for appropriate decisions to be made. Current evaluation of FDI and SHM approaches are based mainly on classification-related performance measures. For this, related numeric and graphical measures are used. Typical evaluation considers a single numeric value describing also the uncertainty of statements. However, contemporary diagnosis systems are concerned with how the characteristics of the fault, change the probability of detecting it. This implies the influence of fault characteristics or process parameter are an essential aspect of the evaluation process. This thesis aims to address this important but often neglected component in the evaluation process.
In the field of Nondestructive Testing (NDT) the ability to detect faults and knowledge of the fault type, location, and dimensions are critical factors in the reliable application of fitness-for-service methods. It is therefore necessary to quantify the effectiveness of NDT methods both to detect and size faults. The concept of Probability of Detection (POD) is used widely in many industries to establish the capability of an inspection procedure. This is generally expressed as a POD curve, which relates the likelihood of detection to a characteristic parameter of the flaw, usually its size. The usefulness and capability of the POD to detect and quantify faults is exploited in this work.
The goal of this research is threefold. The first is to adapt standard POD measurement and implementation strategy to vibration-based health monitoring and therefore integrate vibration analysis results (Results already published in [ARS20], [ARS18b], [ARS18a], and [ARS18c]). The idea is to discuss fault diagnostics from a POD-oriented view as shown in Fig. 1.1.

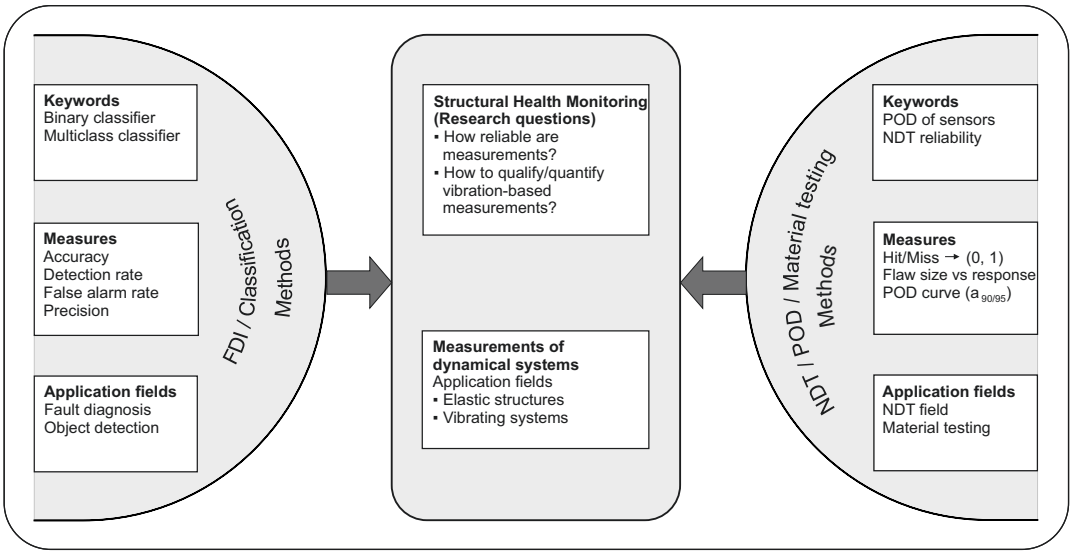

Figure 1.1: Adaptation and implementation strategy [ARS20]

Second is to implement a novel fusion strategy to fuse POD results (Results published in [ARS20]). Conventional FDI/SHM fusion strategies usually consider pre-

cision values as performance measure, which is considered in the fusion process. In this research, it is possible to use measurable POD values as replaceable for the precision value, because both define a performance measure about the reliability of an assignment. Using the fusion approach, the probability of the existence of a fault can be obtained based on the individual performance and assignments of the sensor-/feature-based statements.

Thirdly, the new evaluation approach is extended to other technological fields with different implementation strategies. The purpose of the extension is to investigate the generalization capabilities of the POD approach and also use the proposed approach to address the limitation of the evaluation methods in these fields (Results already published/submitted in [ADS19], [ADS20b] and [BAMS20]).

1.2 Thesis organization

In this thesis a new performance evaluation and fusion methods (obtained in cooperation with Rothe [Rot19] [ARS20]) are developed permitting the integration and assessment of the effect of process parameters on the evaluation results. The thesis comprises of six chapters. Parts of this work have been published/submitted to journals [ARS20], [ADS20b], and [BAMS20] or have been published in proceedings of conferences [ARS18a], [ARS18b], [ARS18c], and [ADS19]. In the current chapter an introduction to the challenges associated with vibration-based FDI/SHM and performance assessment of machine learning and computer vision classification approaches are stated. Open questions and problems are clearly expressed. The second chapter presents a state of the art review of commonly used vibration-based FDI and SHM. Classification-related evaluation approaches are also examined. The existing research gap are identified. Chapter two ends with a new evaluation approach based on the Probability of Detection (POD) measure. The theoretical framework is presented. Chapter three focuses on the adaptation, implementation, and analysis of the newly developed POD in vibration-based SHM. Using a benchmark test rig, the principal difficulty associated with implementation of POD in dynamic systems is demonstrated. A newly defined technique [ARS20] to fuse POD values using probability estimations instead of precision values as performance measure are considered in the fusion process. Using the fusion approach, the probability of the existence of a fault can be obtained based on the individual performance and assignments of the sensor-/feature-based statement. A POD view to fault diagnosis is also introduced. Noise analysis based discussion and a trade-off between decision threshold, Probability of False Positive (PFP) and POD concludes the chapter. In the fourth chapter, evaluation measures like the Receiver Operating Characteristic (ROC) and Precision Recall (PR) curves are examined. These known evaluation curves do not incorporate the effect of process parameter in the evaluation process. The POD is extended to the evaluation of machine learning approaches and incorporates the

effect of process parameter on the classification results. The fifth chapter provides a concept to evaluate classifiers used in computer vision systems for object classification and detection. Explicit and accurate evaluation of CNN classifiers in image detection is examined using the POD. The introduced approach is experimentally evaluated for vision-based classification results of CNN approach considering different image parameters. Finally, the summary of this thesis, conclusions, and future work are outlined in chapter six.

2 Literature review and theoretical background

Before discussing the new diagnostic and evaluation techniques developed in this thesis, it is important to examine current and existing approaches proposed in literature. The purpose is to identify existing research gap and to avoid repeating known knowledge. Structural Health Monitoring (SHM) is relevant in ensuring the safe and optimal operation of elastic mechanical structures. These SHM-systems are used for fault detection, localization, severity quantification and in some cases estimation of remaining useful life. Many statistical models are used to solve detection and localization problems however reliability of diagnosis are often not considered. This chapter presents a state of the art on current Fault Detection and Isolation (FDI) techniques applied to vibrating mechanical systems. A general framework for fault diagnosis is presented. Reliability of diagnostic statements and the effectiveness of approaches is critically examined. Focus and limitations of existing approaches along with the outlook is addressed. The commonly used reliability assessment of Nondestructive testing (NDT) procedure is presented. This NDT reliability measure is based on the Probability of Detection (POD). The theoretical framework of the POD is presented and proposed as a reliability measure albeit different implementation strategies. Improvement of diagnosis techniques by integrating new ideas (filtering, fusion) with classical approaches is proposed.

The content, figures, and tables in this chapter are based on preparation for publication of [AS20].

2.1 Introduction

Elastic mechanical materials are utilized in many engineering structures. Materials that readily come to mind are Aluminum, Carbon Fiber Reinforced Polymers (CFRP) and some grades of steel. The characteristics of these materials that make them attractive to these fields are high strength to weight ratio, resistivity to corrosion, and less susceptibility to brittle fracture. There are also some engineering structures that exhibit elastic behavior. These structures adapt to workload changes so far as the elastic threshold is not exceeded and tend to exhibit high vibrational modes. Notable structures are skyscrapers, bridges, fuselage of aircrafts, turbomachineries, blades and tower of wind turbines. The presence of defects in these elastic systems may affect its structural and functional integrity. From this, it can be concluded that the structural/functional integrity together with the quality of condition monitoring are connected. A fault represent a change or deviation from an acceptable system/process condition [Ise06]. The change could be gradual or abrupt [BN$^+$93] and the extent either local or global [Ise06] [FK09] [CFL03]. Monitoring these deviations require formulation of appropriate diagnostic approaches to

ensure optimal performance and fault diagnosis systems aims to achieve this optimality [KN05] [FW07]. Several SHM procedures exist with vibration-based health monitoring a well-established and utilized technique. Vibration-based SHM use structural properties for fault diagnosis and operates on the principle that structural defects result in changes in dynamical properties. This change in structural properties can be utilized for diagnostic purposes. Fault identification is mainly based on displacement, velocity, or acceleration measurements at a single point [Fri05]; however, useful assessments may include observable effects and multipoint measurements [ARS20] [ARS18a] [ARS18b].
State of health monitoring of elastic structures are complex. The complexity is usually related to three processes: data acquisition, feature extraction, and/or statistical modeling for feature classification [FDN01]. Data acquisition is mainly based on hardware and software systems. Depending on the system monitored, the hardware system could consist of different components. However an important and common component of the hardware system are sensors. Sensors are devices to acquire data for onward processing by other components in the acquisition system. The selection of suitable types of sensors, quantities, and locations need to be effectively addressed. This ensures optimal sensor placement and prevent redundancy.
Feature extraction details the selection and processing of features. This is important due to different sensitivity levels of features to faults. Depending on the application, feature extraction is performed in time, frequency, or time-frequency domain transforming a measured signal to a new representative form for easy and reliable classification. The development of statistical models for fault classification completes the process in many instances. Statistical modeling techniques implement learning algorithms to normalize data and distribution of feature extracted. Operational and environmental effects can conceal damage-related changes. In [FDCS96] [DFG96] methods to distinguish changes caused by defects and environmental conditions are presented.
Structural Health Monitoring (SHM) is the process of implementing a damage detection and characterization strategy for monitoring engineering structures [KN05] [FW07]. Damage in this context describes physical changes that adversely affects the system performance [FW07]. The field of SHM has become an essential aspect of industrial practice to ensure safe operations and improve maintenance [FDN01].
It is noteworthy that the evolution of the damage and changes in the dynamics of the structure act on different time scales [Fri05]. The evolution of the damage is slower compared to the vibration of the structure except for impact damage. The dynamics of vibrating systems changes due to effects like altering the system properties possibly leading to changes in the mass, stiffness, and damping properties. Changes in the structure affects its modal properties. The modal properties can be described by mass M, damping C, and stiffness K by the relation

$$M\ddot{x} + C\dot{x} + Kx = F(t),$$

where x, $\dot{x}$, and $\ddot{x}$ represent the displacement, velocity, and acceleration vectors respectively, with $F(t)$ representing forces and moments. The above mentioned formulation is mainly restricted to the class of linear systems. Transformation into the modal domain yields the eigenvalue equation for the i^{th} mode as

$$(-\omega^2 M + j\bar{\omega}_i C + K)\bar{\phi}_i = 0,$$

where $\bar{\omega}_i$ is the i^{th} complex eigenvalue with it's imaginary part corresponding to the natural frequency ω_i and $\bar{\phi}_i$ is the i^{th} complex mode shape vector with the real part corresponding to the normalized mode shape ϕ_i.

Decisions about the existence of deviations (fault detection) and assignment of causes (diagnosis) have to be made. Several FDI techniques are currently utilized to detect and isolate faults. These methods can be primarily grouped into four categories: signal-based, model-based, data driven, and hybrid approaches. Signal-based approaches utilize output signals. Fault diagnosis models compare raw or filtered signals to thresholds and conclude to the presence of faults. Model-based approaches use beside output signals, input signals and requires a model to be built (parameter identification), to be assumed (observer), or to be suitably established [SWWS16] [IB97]. Data-driven approaches formulate implicit relationships by trained models through analysis of fault-free data obtained during regular operations [IB97]. These models are used to estimate the behavior of variables to be compared with those obtained from measurements [MJ11] [DG13] [FK09] [Qin12] [RJS15]. Hybrid approaches combine model-based and model-free techniques [DG13] [LMSM$^+$16] [RJS15].

Vibration-based feature extraction can be performed in time, frequency or time-frequency domain of the signal [ABSC11] [Can10]. In time domain representation, the data are displayed as a function of time. Fault diagnosis in the time domain usually utilize statistical techniques [FS09] [GC09]. Time series methods are common in vibration fault analysis [KF10] [CB08] [KF13]. In the frequency domain, the Fourier transform (FT) is the most basic and appropriate for sparse representation of a signal [Che15]. The FT maps a signal into its frequency components [Che15]. Time-frequency signal processing is a set of signal processing methods, techniques, and algorithms in which the two variables time and frequency are used concurrently [ABSC11] [Fug09] [Mal99] [Sta10] [DSJ08] [AB12]. This technique deviates from the classical methodologies in which time or frequency representation is used independent of the other [BLW$^+$87] [SHZ$^+$14] [HSL$^+$98]. Short Time Fourier Transform (STFT) is a widely used time-frequency extraction tool due to its low computational costs in comparison to other time-frequency signal processing tools [MS08] [LW98] [WD99]. Vibration-based feature extraction methods in specific domains are shown in Tab. 2.1.

Table 2.1: Vibration-based feature extraction techniques [Mar12]

Time	Frequency	Time-Frequency
Fractal analysis	HFRT	Wavelet packet trans.
Shock pulse method	Envelope Analysis	WVD
Time series averaging	Cepstrum analysis	WT
Statistical features	Adaptive noise cancellation	MFCCs
RMS	Spectral analysis	EMD
Peak value	Freq. averaging techniques	Spectral Kurtosis
Skewness	PSD	CWT
Kurtosis	Statistical features	DWT
Crest factor	FC frequency center	STFT

2.2 FDI approaches applied to elastic mechanical systems

This section details vibration-based FDI approaches implemented in the fault diagnosis of elastic structures. These methods can be primarily grouped into frequency based-methods, mode shape based-methods, curvature mode shape-based methods, and methods that combine mode shapes and frequency. There are contemporary methods based mainly on machine learning approaches and these are classified differently.

2.2.1 Frequency based-methods

There are sizable literature related to fault detection using shifts in resonant frequencies. The knowledge that changes in physical properties results in changes in resonant frequencies contributed significantly to the use of this method for damage identification. A very detailed review and reference list can be found in Doebling [DFPS96]. The use of frequency shifts have practical limitations in many applications because it is difficult to distinguish frequency shifts caused by faults and those produced by changes in environmental conditions. The somewhat low sensitivity of frequency shifts to damage requires either very precise measurements or large levels of damage [DFP$^+$98]. Studies in the late 1990s [FDCS96] [DFG96] have depicted that resonant frequencies have less statistical variation from random error sources in comparison to other modal parameters. For example, in offshore platforms damage-induced frequency shifts are difficult to distinguish from shifts resulting from increased mass from marine growth [DFP$^+$98]. In [PGO14], vibration testing of composite laminates to detect impact damage is presented. Drop-weight impact experiments are carried out on 44 carbon fibre reinforced composite laminated specimens with incident energy levels ranging from 6.6 to 70 J. This is followed by compression after impact (CAI) test to determine the compressive residual

strength of the specimen. Frequency shift is used as an indicator for interlaminar damage onset (IDO) and barely visible impact damage (BVID). The significant variations is observed in the 4th, 8th, 9th, 10th and 11th transverse modes and in the 23rd and 24th in-plane modes. The other damage indices used are the mode shape changes, curvature mode shape changes and residual load-bearing capacity. Detection of BVID in aircraft composite structures based on nonlinear elastic material behaviour of damaged materials is addressed in [PM09]. The specimen used for the experiment are carbon fiber composite plates. Damage is introduced into the plates by low velocity impact. Two damage indices: a)resonant frequency shift b)presence of harmonics and sidebands, are used. This is based on the principle that the presence of damage results in phenomena like hysteresis. The damaged material can be described by the relation

$$\sigma = \int K(\varepsilon, \dot{\varepsilon}) d\varepsilon,$$

where $K(\varepsilon, \dot{\varepsilon})$ is the nonlinear modulus. The hysteric modulus is given by

$$K(\varepsilon, \dot{\varepsilon}) = K_o(1 - \beta\varepsilon - \delta\varepsilon^2 - \alpha[\Delta\varepsilon + \epsilon(t)sign(\dot{\varepsilon}) + ...]),$$

where K_0 is the linear modulus, $\Delta\varepsilon$ is the strain change amplitude, α is the material hysteresis measure, β and δ nonlinear coefficients to be determined. The hysterical behavior of the damaged material is analyzed by correlating the third harmonic amplitude with the fundamental amplitude. The authors however proposed the validation of the approach for different materials and lay-up configuration.

2.2.2 Mode shape based-methods

West [Wes86] introduced the use of mode shape information for structural damage localization without prior FEM usage. The author uses the Modal Assurance Criteria (MAC) to determine the correlation between modes from the test of an undamaged space shuttle orbiter body flap and the modes from the flap after it has been exposed to acoustic loading. The change in MAC across different partitioning techniques is used to localize the defect. Fox [Fox92] demonstrate that mode shape changes using single number such as MAC are relatively insensitive to saw cut damage in a beam. A MAC approach based on measurement points close to a node for a particular mode, is found to be a more sensitive indicator of changes in the mode shape. When examining eigenfrequency, graphical comparisons of relative changes in mode shapes proved to be a better way of detecting fault location. In [May91] a method for model error localization based on mode shape changes is presented. The method is based on structural translational and rotational error checking (STRECH). By taking ratios of relative modal displacements, STRECH assess the accuracy of the structural stiffness between two different structural degrees of freedom (DOF). The STRECH approach can be applied to compare the results of

a test with an original FEM or to compare the results of two tests. In [Liu13], modal analysis of the complete blade-cabin-tower system is presented. The coupling of the entire blade-cabin-tower system by developing coordinate systems and kinetic equations are established as a tool for vibration fault diagnosis. The natural frequencies of the individual components are calculated based on the coordinate system and the random wind vibration analyzed. The introduced approach provides a means to analyze the complete system without individual component analysis. In [DSP+16], two methods: a) coordinate modal assurance criterion (COMAC) and b) modal strain energy method for the detection of damage in a composite helicopter main rotor blade are compared. The PZL SW-3 Sokol helicopter blade is experimented on. The blade spans approximately 7 m with a varying cross section. To obtain a free-free boundary condition, the blade is suspended with two elastic cords at the ends. Acceleration measurements are taken at 55 points using 55 accelerometers. Beam-wise and torsional dynamics are captured with 49 accelerometers and the in plane dynamics measured by 6 accelerometers. An electrodynamic shaker is used as a source of excitation and the pristine blade accelerations (baseline data) measured by the accelerometers. A 300 g mass is attached on the blade to simulate mass unbalance and a second experiment undertaken. PolyMAX frequency domain identification method is used to identify the vibration modes of the blade. From the results, modal strain method is more sensitive and precise in localizing the mass but noisier compared to the COMAC. The authors propose the use of strain gages or Fiber Bragg Grating sensors, to directly measure strain modes for onward strain energy formulation.

2.2.3 Curvature mode shape-based methods

The use of mode shape derivatives, such as curvature is an alternative to obtain information about structural damage. For elastic structures such as beams,plates, and shells, there is a direct relationship between curvature and bending strain [DFP+98]. Pandey et al. [PBS91] show that absolute changes in mode shape curvature is a good indicator of damage for FEM beam models considered. The curvature mode values are calculated from the displacement mode shape using the central difference operator. In [SKT92], the authors present a method based on the decrease in modal strain energy between two structural DOF, as defined by the curvature of the measured mode shapes. Topole and Stubbs [TS95a] [TS95b] examine the feasibility of using a limited set of modal parameters for structural damage detection. In [SK96] the authors examine the feasibility of localizing damage using curvature mode shape values without baseline modal parameters. Chance et al. [CTW94] investigated and concluded that numerical computation of curvature from mode shapes resulted in unacceptable errors. They used measured strains instead to measure curvature directly, which improved the results.

2.2.4 Methods that combine mode shapes and frequency

Using single features for diagnosis purposes may not be effective in some applications. To overcome this limitation, hybrid approaches are implemented. One such hybrid approach is the combination of frequency and mode shapes. Utilization of modal analysis to detect fatigue damage in Aluminum cantilever beam is presented in [RKP11]. The natural frequency and mode shapes (measured or modeled) from frequency response function is used as damage indicator. It is shown that the proposed technique is suitable for damage localization in beam-like structures.

In [YS13], a review on testing, inspection and monitoring approaches for wind turbine blades is presented. The survey indicate that three mechanical property testing procedures exist: static testing, fatigue testing, and modal testing. In static testing, loads are applied statically to determine the ultimate strength of the blade. This is done to examine the blades ability to withstand torrential winds. In fatigue testing is performed to investigate structural properties, fatigue strength, and failure modes. Modal testing characterize the damping and dynamic properties like natural frequency and mode shape. Another category, full scale testing, incorporates all three techniques. The authors predict full scale testing will become the most significant performance validation method. Also real-time, remote, wireless, and smart SHM systems are predicted to play an important role.

2.2.5 Machine learning approaches

Most contemporary fault diagnosis approaches are based on machine learning methods. Application of machine learning methods can be found in [JVW10] [SSB10]. In [TCO08], modified PCA and pattern recognition algorithm is used in the detection of faults in a scaled finite element Aluminum alloy model of an aircraft wing. The wing structure is divided into five sections. Two forms of faults are considered: cracks and distributed damage (introduced into a section by stiffness reduction). The modified PCA reduces the dimensions of the frequency response function and projects the data into a plane that makes the classification task easier. The 1-NN classifier based on the Euclidean distance between two feature vectors is used to classify the feature as damaged or not. The authors suggest the methodology is suitable for structural damage detection but proposed the implementation of the approach on experimental data.

In [BS15], a review on the non-linearities in the vibrations of elastic structures induced by breathing crack is presented. The structures focused on are shafts, beams, and plates. The paper suggests advantages of using non-linear vibration effects in the case of shafts and beams because of the high sensitivity to the effect of breathing crack, but same can not be said for plate-like structures. Since cracks in metallic plates are mostly through-the-thickness cracks and does not constitute

breathing cracks. It is concluded that none of the reviewed methods has universal applicability and that a method is selected based on the fault type. The cons in using these approaches are the high modeling complexity and exhausting calculation. The authors propose mathematical signal processing like wavelets, novelty detection approaches among others.

Despite the advances in vibration-based SHM applied to elastic mechanical structures, transition from theory to practical implementation, early detection of faults, quantification of fault severity, and reliability assessment of diagnostic statements remains an open problem [SWWS16] [KN05].

2.3 Evaluation of FDI approaches

Performance evaluation and analysis of FDI approaches are mainly based on classification related performance measures. These assessments are related to either a numeric measure or a graphical representation. The numeric measures are based on the calculation of True Positives (TPs), False Positives (FPs), False Negatives (FNs), and True Negatives (TNs). They can be illustrated in a confusion matrix (Fig. 2.1) as a basic performance metric [Sus04]. Advanced evaluation metrics can

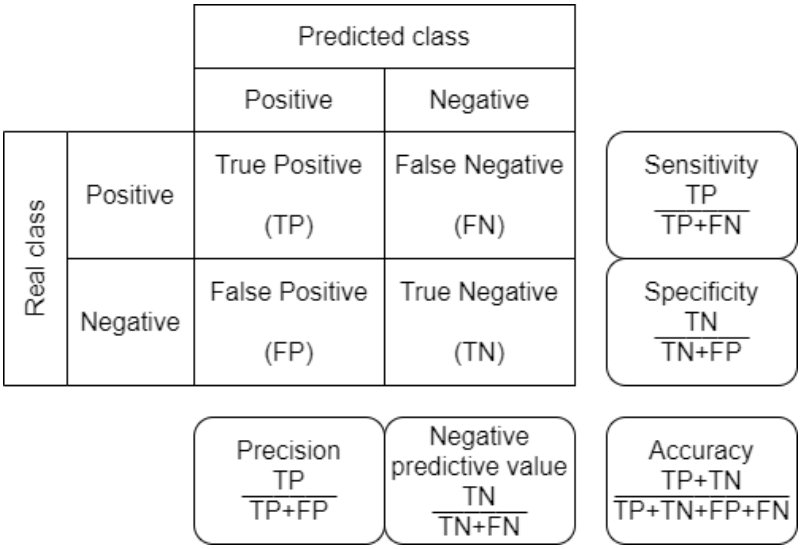

Figure 2.1: Existing evaluation approaches based on confusion matrix

be calculated from the basic numeric measures. These advanced metrics include sensitivity, negative predictive value, positive predictive value, precision, specificity, accuracy among others. Selection of a suitable numeric measure normally depends on the classification purpose.

There exist many graphical assessment techniques with the Receiver Operating Characteristic (ROC) and Precision-recall (PR) curve the predominantly used. The

ROC is a graph indicating the performance of a classification model as its discrimination threshold is varied. The ROC is constructed with Detection Rate (DR) against the False Alarm Rate (FAR). The ROC is a probabilistic curve and the area under the ROC (AUC) represents degree of separability. The AUC tells how much a model is capable of distinguishing between classes. The PR curve on the other hand is a plot of the precision and the recall for different probability thresholds. The PR curve is recommended for imbalance class data. Despite the popularity of both curves, the influence of flaw characteristics on the classification results are not considered. This thesis attempts to solve this limitation using the Probability of Detection reliability measure.

The principal aim of POD 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 combined in the POD approach. The POD development and implementation strategies are outlined in the next section.

2.3.1 Probability of Detection

The introduction of damage tolerance decision rules as a way to determine the integrity of a component resulted in the institution of statistical tools for assessing the detection probabilities for NDT systems [SUS13]. The first general requirements to quantify the capabilities of NDT was established with the design and production of National Aeronautics and Space Administration (NASA) space shuttle system in 1969 [Geo07]. The methods and designs introduced by NASA were soon adopted by the United States Air Force (USAF) in the early 1970s [PH]. The background and evolution of the POD are indicated in Tab. 2.2.

Table 2.2: Evolution of the POD measure

Approach	Property	Developer	Year
Moving average	Hit/miss	NASA	1969
Binomial stats	Hit/miss	Yee et al.	1976
Log. regression	Hit/miss	USAF	1978
Cum. distribution	Hit/miss	Berens & Havey	1988
Automated EC	Hit/miss Target response	Packman et al.	19991
Probit, Logit	Hit/Miss Target response	Berens	1993
MAPOD	Hit/miss	Thompson & Meeker	1997
Cum. distribution w/ imp. stab. bounds	Hit/Miss Target response	DOD: MIL-HDBK-1823A	2009

Since these pioneering groundworks, extensive research activity and experimental programs have followed over the course of the years. Probability of Detection is now an established certification tool [Geo07]. Data used in producing POD curves are categorized by the main POD controlling factors/variables. These factors/variables are either discrete or continuous and can be classified as [MHA09] [Geo07]

1. Target response: systems which provide quantitative measure of target and
2. Hit/miss: produce binary statement or qualitative information about the existence of a target.

The theoretical framework for both approaches are illustrated below.

2.3.2 Target response approach to POD

The target response approach is used when there exist a relationship between a dependent function and an independent variable [MHA09]. In the derivation of the POD curve, a predictive modeling technique is required. One such method is regression analysis of the data gathered [MHA09] [Ann17] [GA10]. The data distribution could be linear or not. A strategy to linearize the data distribution is by plotting four models: *X vs Y, log X vs Y, log Y vs X, and log X vs log Y.* The model with best linearity and variance is used in the construction of the POD curve [KNN$^+$05]. The regression equation for a line of best fit to a given data set is given by

$$y = b + mx, \tag{2.1}$$

where m is the slope and b the intercept. The Wald confidence bounds are constructed to define a confidence interval that contains 95 % of the observed data [KNN$^+$05]. Here the 95 % Wald confidence bounds on y is constructed by

$$y_{a=0.95} = y + 1.645\tau_y, \tag{2.2}$$

where 1.645 is the 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 [MHA09] [Ann17]. 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

$$z_i = \frac{(y_i - (b + mx_i))}{\tau} \tag{2.3}$$

and

$$f_i = \prod_{i=1}^{n} \frac{1}{2\pi} e^{-\frac{1}{2}(z_i)^2} \tag{2.4}$$

the information matrix I can be computed as

$$I_{ij} = -E(\frac{\partial}{\partial\Theta_i \partial\Theta_j} log(f)). \tag{2.5}$$

The inverse of the information matrix yields ϕ as

$$\phi = I^{-1} = \begin{bmatrix} \sigma_b^2 & \sigma_b\sigma_m & \sigma_b\sigma_\tau \\ \sigma_m\sigma_b & \sigma_m^2 & \sigma_m\sigma_\tau \\ \sigma_\tau\sigma_b & \sigma_\tau\sigma_m & \sigma_\tau^2 \end{bmatrix}. \tag{2.6}$$

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

$$\Phi(\mu, \sigma) = \frac{1}{2}\left[1 + erf\frac{x-\mu}{\sqrt{2}\sigma}\right]. \tag{2.7}$$

The POD as function of target a is derived as

$$POD(a) = \Phi\left[\frac{a-\mu}{\sigma}\right]. \tag{2.8}$$

2.3.3 Hit/miss approach to POD

An efficient implementation of the binary data is to posit an underlying mathematical relation between POD and parameter and consequently model the probability distribution [MHA09]. The use of ordinary linear regression is inappropriate because the data are not continues but discrete and bounded. Generalized Linear Models (GLM) overcome this challenge by linking the binary response to the explanatory variables through the probability of either outcome, which does vary continuously from 0 to 1 [NW72] [MHA09]. The GLM attains this through [NW72]

1. A random component specifying the conditional distribution of the response variables, Y_i (for the $i-th$ of n independently sample observations)
2. A linear predictor that is a function of regressors
3. A smooth and invertible linearizing link function $g(y)$, which transforms the expectation of the response variables $P_i \equiv E(Y_i)$ to the linear predictor.

The transformed probability can then be modeled as an ordinary polynomial function, linear in the explanatory variables. The POD can be generated from the GLM as explained in the case of linear regression. The commonly used GLM in POD are the log, logit, probit, loglog, and weibull [MHA09]. Depending on the data distribution a model may be appropriate compared to the other. One criteria used is to select GLM with least deviance.

The department of defense (USA) and many authors/institutions consider MIL-HDBK-1823A as the state of the art and contemporary guide for POD studies [SKGDD15] [MQLDM18]. The algorithms and programs of MIL-HDBK-1823A [Ann17] [MHA09] will be adapted in this thesis. The POD approach is proposed as a diagnostic and reliability measure to be implemented in vibration-based FDI and other technological fields.

2.4 Summary and conclusions

In this chapter, existing approaches used in vibration-based FDI are presented. These methods are grouped into frequency based-methods, mode shape based-methods, curvature mode shape-based methods and methods that combine mode shapes and frequency. Contemporary methods are based mainly on machine learning approaches. Several FDI approaches are utilized however reliability evaluation of FDI schemes are usually not considered. Also, the ability of existing approaches to quantify faults, transition from theory to practical implementation, and a standardized certification standard remains a big challenge. In comparison to NDT approaches, the POD is well known and utilized. The POD serves two purposes. It can be used as a fault detection and quantification method. Secondly, it can be used as a reliability measure and a certification tool. Adapting and implementing the POD approach in vibration-based FDI and as a reliability measure are proposed and subsequently used in this work.

3 Fault diagnosis and damage quantification of elastic mechanical systems

In this chapter, fault diagnosis of elastic mechanical systems is presented. The complexity associated with monitoring vibrating systems is introduced. Besides challenges associated with dynamical behaviors of the systems monitored, supervision tasks are complex with respect to data acquisition, feature extraction, and/or statistical modeling for feature classification. A diagnostic technique to access the mentioned challenges is developed. A new fusion approach is introduced to improve the detection quality by using suitable assumptions in combination with sensor/information fusion applied to feature-based analysis as detection task. A new noise analysis procedure is introduced permitting the selection of a decision threshold with the corresponding detectable flaw size and related false alarm rate. Selecting different sensors implies changing the signal distribution character and the decision threshold. This change results in different values and hence can be exploited to decide the optimal sensor. The new approach provides a graphical representation that illustrates the diagnostic capabilities of a sensor as its decision threshold is varied.

The content, figures, and tables in this chapter are based on publication of [ARS20], [ARS18a], [ARS18b], and [ARS18c].

3.1 Introduction

Elastic mechanical systems refer to structures that adapt to workload changes provided the elastic limit is not exceeded. Notable structures are skyscrapers, bridges, aircrafts, wind turbines among others. These structures exhibit higher vibration modes when excited and are also complex, safety-critical, and capital intensive. Materials like aluminum, carbon fiber reinforced polymers (CFRP), and some grades of steel exhibit these elastic characteristics and are commonly used in applications requiring high strength-to-weight ratio and corrosion resistance. Ensuring structural and functional safety guarantee optimal performance. However the presence of defects in components may affect its structural and functional integrity. From this, it can be concluded that the structural/functional integrity and the quality of detection and diagnosis approaches are connected. During the last decades, several FDI techniques are implemented to detect changes, faults, and local defects. These approaches can be categorized into four groups: signal-based, model-based, data driven, and hybrid approaches. Signal-based approaches use output signals. Fault detection modules often compare raw or filtered signals to thresholds and conclude to the presence of faults. Data-driven approaches establish relationships using trained models through analysis of fault-free data obtained during normal operations [IB97]. Hybrid approaches combine model-based and model-free techniques [PJK+16].

Signal-based methods are easy to implement and widely adopted in fault diagnosis. The main idea is to extract relevant process characteristics from analyzed sensor data. These characteristics are subsequently combined with further knowledge related to specific health states of the system [ASSS14].
Evaluation assessment of FDI approaches are based on classification-related performance measures (PM). These PMs are numeric values or graphical representations. For this, related measures like DR or FAR are evaluated based on the confusion matrix. When discussing the effects of varying tuning parameters the complete precision values of the ROC can be used for evaluation [WY07]. Typical evaluation considers a single set of numbers like accuracy which is used in describing the uncertainty of statements. In some applications fault tolerant control systems apply FDI techniques and reconfigure controllers to enhance the entire system reliability [WLL$^+$13] [HZY07]. Reliability evaluation of FDI schemes itself still remains an open problem. The overall aim of reliability assessment is to guarantee optimal performance and SHM aims to achieve this optimality.
Structural health monitoring is the process of implementing a damage detection and characterization strategy for monitoring engineering structures [KN05] [FW07]. Damage in this context describes physical changes that adversely affects system performance [FW07]. The field of SHM increasingly has become an essential aspect of industrial practice to ensure the quality of products, safe operations, improved maintenance, and to save cost. 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]. Nondestructive 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. Nondestructive testing consists of a variety of non-invasive inspection techniques used to evaluate materials, components, structures or entire process units.

Despite the technological advancements in the field of SHM, salient questions persist. These include the transition from developed theory to practical implementation, early detection of faults, and reliability assessment of diagnostic statements [SWWS16] [KN05]. The dynamics of vibrating systems changes due to different effects like altering of the structure possibly leading to changes in the mass, stiffness, and damping properties. Finally decisions about the existence of changes (fault detection) and/or specific faults (diagnosis) have to be made. Making reliable decisions is of importance. Current FDI approaches utilize classification-related performance measures. These performance measures are subject to uncertainties hence not quantified. No clear or defined reliability standard also exist. Conversely, conventional NDT approaches use the Probability of Detection as a reliability measure. The emergence of SHM led analysts to ponder how to integrate POD in SHM systems. In comparison to NDT, SHM systems are mounted permanently and should provide reproducible results. Aging of the structure leads to adverse effects. In Aldrin et

al. [AASL16] Monte Carlo simulation of flaw size as a function of time is proposed. The results demonstrate the sensitivity of flaws to degradation of SHM-system. In NDT, uncertainties are associated with human factors, variations in the interface between the structure and transducer, crack form and shape, and local structure properties whilst SHM uncertainties are environmental conditions, aging effects in the structure, and damage morphology [SUS13]. In Mendrok and Uhl [MU10], experiments are undertaken on an Aluminum frame to ascertain the POD capabilities of a modal filter-based damage detection system. For monitoring purposes, ten accelerometers are mounted on different positions on the frame. An impulse test is carried on the specimen and the obtained damage index values and the corresponding damage sizes are the input data to determine POD curve. However, no suitable fusion technique for the individual sensor data is developed. Also, variations in fault position and severity are not considered. Resource and time constraints in POD analyses has given rise to Model Assisted POD (MAPOD). Model Assisted POD is proposed in Aldrin et al. [AML$^+$11] and Mueller et al. [MJB$^+$11] utilizing numerical models; however, numerical efforts and computational time difficulties have to be solved for convenient applications in practice. Also, the ability of these models to simulate real faults remains a challenge. The POD approach typically quantifies a sensor/filtering technique in combination with mostly static measurements. Implementing the POD in the field of vibration-based SHM is difficult. Difficulties result from the complexity of dynamical behavior in relation to faults, sensors position (observability), related data analysis procedures, and the system under consideration. The POD approach usually uses a so-called POD curve, constructed by plotting the accrual of flaws detected against the flaw size or produce a response over a threshold [Geo07] [MHA09].

3.2 Implementation of POD in fault diagnosis

According to [SUS13] [Geo07] [MHA09] the POD approach allows a general assessment of the reliability of NDT methods and lately SHM-systems. The aim of the $a_{90/95}$ criteria is to specify a damage size, which can be detected/missed applying a specific method to be evaluated, taking into account statistical variability of the sensor and measurement properties. The United States Air Force (USAF), National Aeronautics and Space Administration (NASA), as well as many authors and institutions consider MIL-HDBK-1823A (updated version of MIL-HDBK-1823) as the state of the art and contemporary guide for POD studies [SKGDD15] [MQLDM18]. Unlike traditional POD implementation strategies, this paper adapts the method but implements a filtering approach to extract features and use as response. It is worth mentioning that this research focuses on fault detection. Here classical sensors in combination with a vibration-oriented system response analysis are utilized. Difficulties associated with observable effects, faults and sensor positions, and the

non-uniqueness of the POD curve characteristics for which not so much attention is given in literature with respect to these combinations are demonstrated.

3.2.1 Adaptation and application of POD in vibration-based fault diagnosis

The aim of this research is to adapt standard POD approaches to vibration-based health monitoring and therefore integrate vibration analysis results (as shown in Fig. 1.1). Conventional NDT approaches apply statistical analysis and show the relationship between signal strength $\hat{a}$ and the size a of the flaw producing the response. The time and cost involved in POD analysis has given rise to Model-Assisted POD (MAPOD) to improve the effectiveness of POD models with little or no specimen testing by utilizing model generated data [KALA07]. Two MAPOD methods are known [TBF+09]. The first utilizes physics-based models to propagate directly the uncertainty of a given set of examination parameters. Second is a transfer function approach which is also a physics-based method that transfers the computed POD curve characteristics for a specific process to another with different parameters [HHB09].

Since SHM systems are permanently mounted, data are continuously recorded. Through signal processing, suitable features (e.g. eigenfrequencies and band power) are extracted from time series data. The feature extraction task reduces data randomness and noise effects. Uncertainty in measured data are accounted by constructing prediction bounds. The prediction bounds ensures that for every new 100 observations, 95 of them should fall within the constructed prediction bounds. Currently, classification approaches are used in monitoring vibrating systems (illustrated in Fig. 1.1) but incapable of quantifying vibration-based fault measurements. However, the POD measure from the NDT and material testing field can be effectively implemented in vibration-based fault diagnosis to quantify sensor-related measurements and therefore be used alongside known classification approaches.

3.3 Experimental results

Illustration of the complexity related to data acquisition, feature extraction, and statistical modeling for feature classification and its effect on the probability of detecting faults is presented in this section. Using a benchmark test rig the principal problem associated with implementation of POD in vibration-based SHM is demonstrated. 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 obtained results and the analysis are discussed in detail.

3.3.1 Experimental set-up

The experiment is carried out on an elastic mechanical beam using the test rig in Fig. 3.1. An elastic steel beam of dimensions 545x30x5 mm is clamped on one side. The beam length is divided into five equal parts (Fig. 3.2) defining sensors position. Two strain gauges 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. Piezoelectric accelerometers are attached at three positions (P1, P2, and P3) on the beam. The beam can be excited by modal hammer.

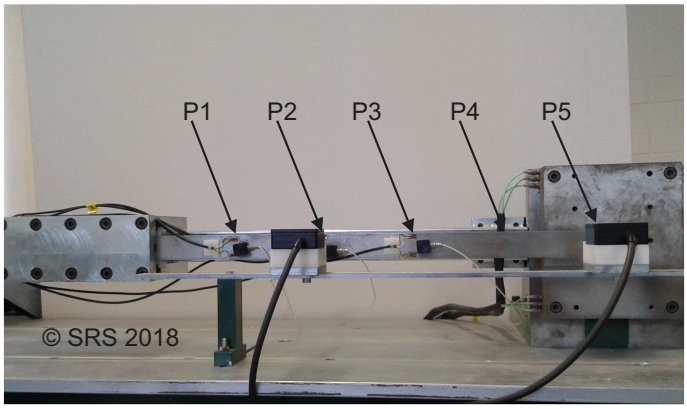

Figure 3.1: Test rig (Chair Srs, UDuE) consisting of a one side clamped elastic beam with bonded strain gauges, laser sensors, and accelerometers [ARS18b]

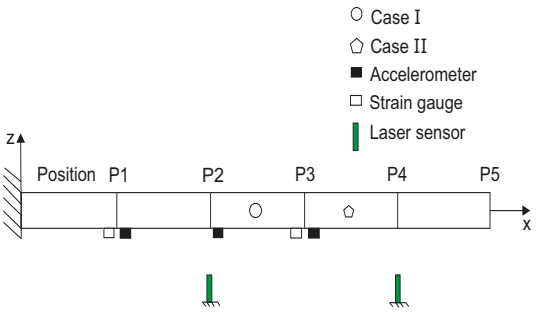

Figure 3.2: Sensor positions relative to beam length [ARS18c]

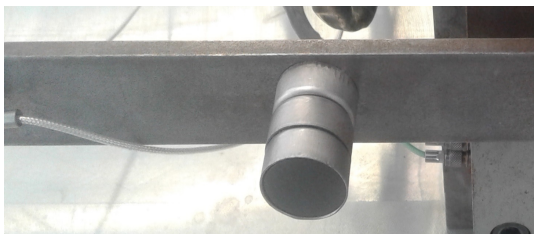

Figure 3.3: Mechanical beam modification using additive mass [ARS20]

3.3.2 Injected faults as changes to be investigated

In this paper changes within the elastic mechanical structures are assumed as changes due to varying mass, so here additive masses (illustrated in Fig. 3.3) are applied to modify the existing initial system to simulate a fault. Two cases of point mass placement are examined (illustrated in Fig. 3.2). 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 recorded.

3.3.3 Results

The analysis is carried out for frequencies associated to the first and second mode for each situation of mass placement (cases I and II). In Fig. 3.4 time series data for different sensors used in the experiment are given. The regression analysis, confidence bounds, prediction bounds, and the POD characterization for each sensor are carried out.

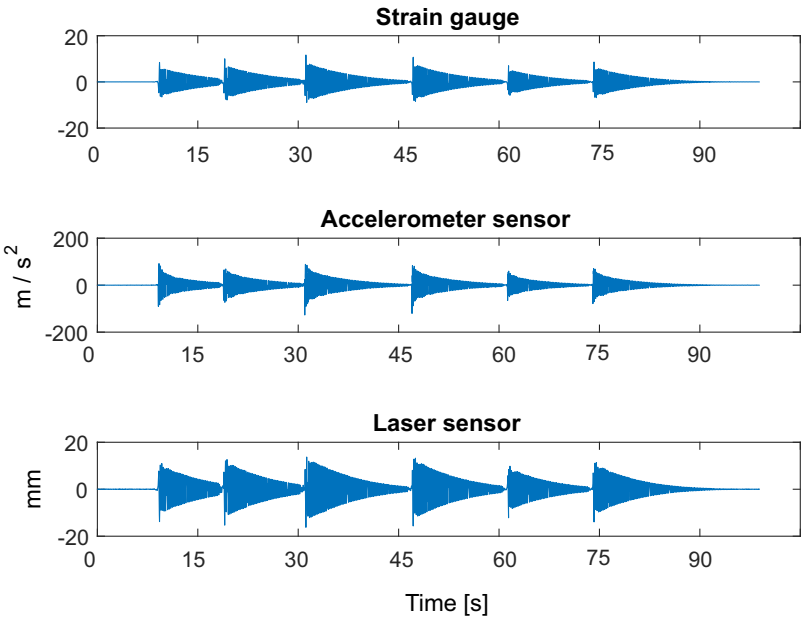

Figure 3.4: Signals of different sensors [ARS20]

The strategy to map the data to the POD curve is shown (Fig. 3.5). In Fig. 3.5 a graphical representation of the flaw size vs. response approach elaborated earlier is given. It involves setting threshold values and fitting trendline to the data. Confidence and prediction bounds are constructed on both sides of the line of best fit. Probability density functions at each flaw size are established. The area above the decision threshold is used to construct the POD curve.

The introduced strategy is implemented to analyze the frequency results from all sensors. The results are given in Tab. 3.1.

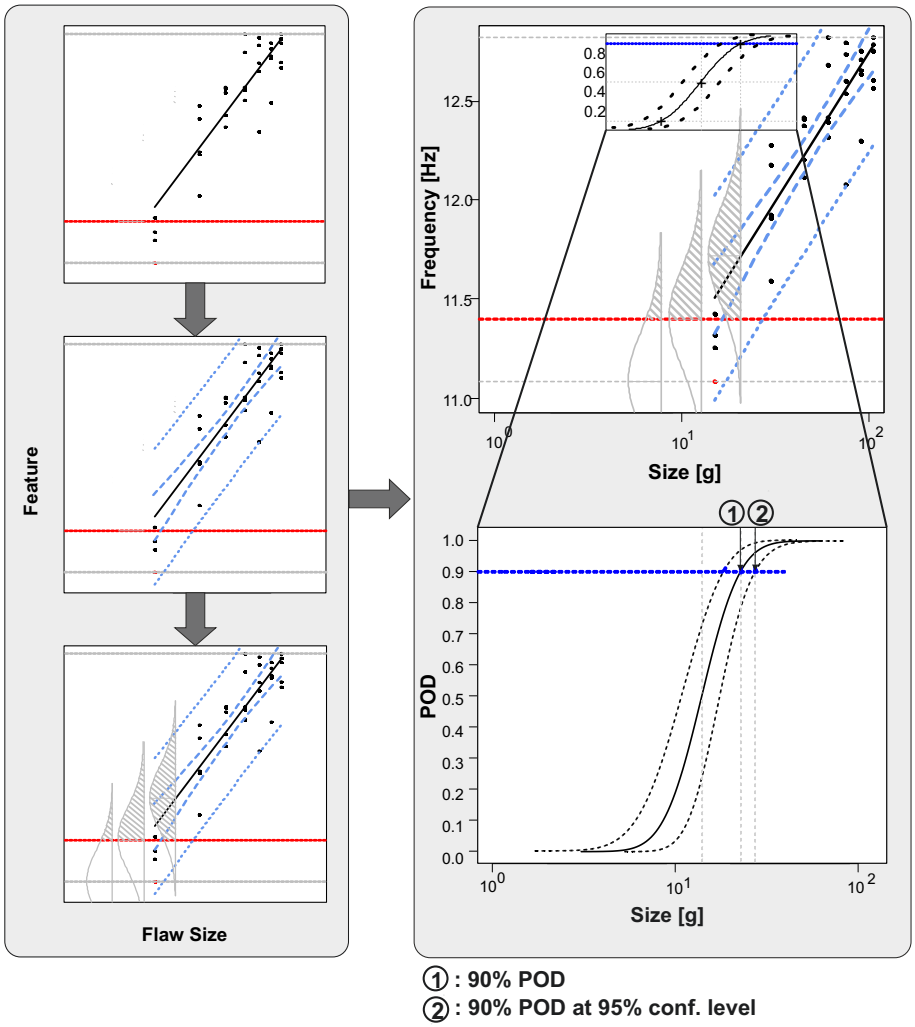

Figure 3.5: Strategy from data to POD curve [ARS20]

Table 3.1: Measure: Eigenfrequencies (Mode 1 and 2) and band power [ARS20]

	Point mass between P2 and P3 (Case I)			Point mass between P3 and P4 (Case II)		
Sensor	Mode 1 freq. $a_{90/95}$ POD (g)	Mode 2 freq. $a_{90/95}$ POD (g)	Band power $a_{90/95}$ POD (g)	Mode 1 freq. $a_{90/95}$ POD (g)	Mode 2 freq. $a_{90/95}$ POD (g)	Band power $a_{90/95}$ POD (g)
ACC 1 at P1	74.04	48.15*	45.28**	52.21	9.915	84.56**
ACC 2 at P2	74.04	55.78	34.63	52.15	9.293*	22.78
ACC 3 at P3	74.04	72.59**	20.20*	52.15	13.36	17.29*
SG 1 at P1	85.19	72.38	29.34	52.15	11.07	28.69
SG 2 at P3	126.70**	62.23	34.37	54.08**	9.394	27.93
Laser 1 at P1	67.30*	61.03	27.64	52.10*	43.49**	25.54
Laser 2 at P4	74.04	-	23.44	52.15	-	25.85

Legend: ACC: Accelerometer, SG: Strain gauge, P: Position, *: Best results, **: Worst results

From the results, it becomes evident that the $a_{90/95}$ POD quantification is different depending on sensor type and mode considered. Based on fault position, sensor

location relative to fault, sensor type, and the mode considered, different results are obtained. The numbers indicate the maximum mass that can be missed with a 90 % POD at a 95 % confidence level. The lowest masses (*) represent best results, so the sensor detects least mass change. The worst (**) POD sensor characterization, represent worst results so the sensor requires large fault values to be detected with $a_{90/95}$ reliability. To explore the non-uniqueness of a feature for POD analysis, as additional feature, band power is also extracted. The band power represents the average power in the frequency range (here: 0-500 Hz). Analysis for this feature is given in (Fig. 3.6 and 3.7). The results are compared with the eigenfrequency results (Tab. 3.1).

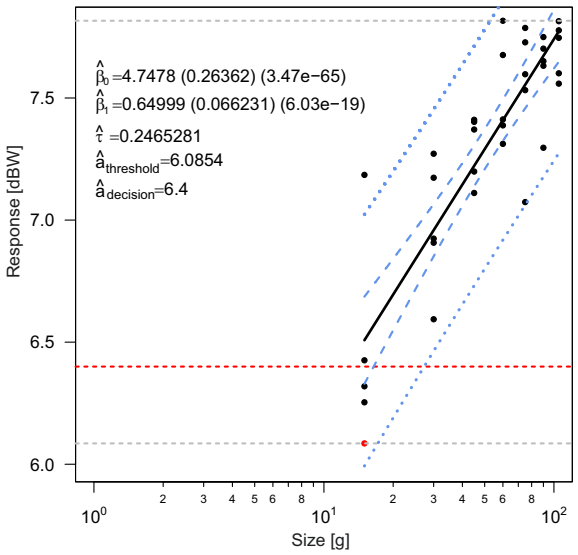

Figure 3.6: Regression analysis related to band power [ARS20]

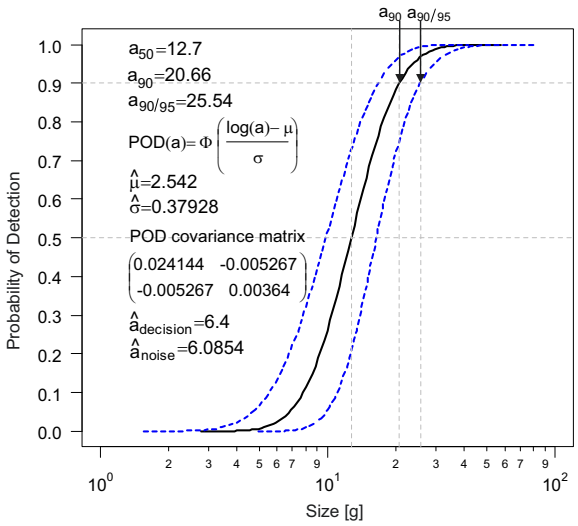

Figure 3.7: POD related to band power feature [ARS20]

It can be concluded that the POD of vibration-based analysis of elastic structures strongly depends on sensor type (dynamic range differs for each sensor), sensor position (same sensor at different positions produce different results), fault position (case I or II), and attribute selected (eigenfrequency or band power). Consequently it can be stated that general statements about the usefulness of related sensors are not possible. Depending on effects, measurement options, features considered, and sensor type the choice becomes sophisticated and task-specific.

3.4 Fusion of results to combine individual FDI-statements

This section introduces a new fusion strategy for POD values. The precision value is an influencing factor to the overall accuracy of fused results. In some applications, the precision value is not computable due to unavailability of training or validation data. Probability estimations are available as classifier output denoting the likelihood of a sample belonging to the assigned class. This value can directly replace the precision value, because both represent probabilities belonging to one class. Furthermore, conventional NDT approaches use the POD as a reliability measure instead of the precision value. The POD is parameter dependent and quantifies a sensor-filtering approach with respect to mostly static measurements, so the replacement of the precision values with the POD values is more complex. This section details a new concept replacing the precision values with POD values. This is based on utilization of the experimental results by Rothe [ARS20] [Rot19].

3.4.1 Fusion concept

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-feature combination for specific flaws, (here: the $a_{90/95}$ POD values for the sensor-feature combination) can be used to calculate the belief values according to the Bayesian combination rule. The procedure is shown in Fig. 3.8. 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 ($POD_j(a_{90/95}, i)$ 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 Pj used for fusion is set as $P_j(a_{90/95}, i) = POD_j(a_{90/95}, i)$ in case of detection and $P_j(a_{90/95}, i) = 1 - POD_j(a_{90/95}, i)$ 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.

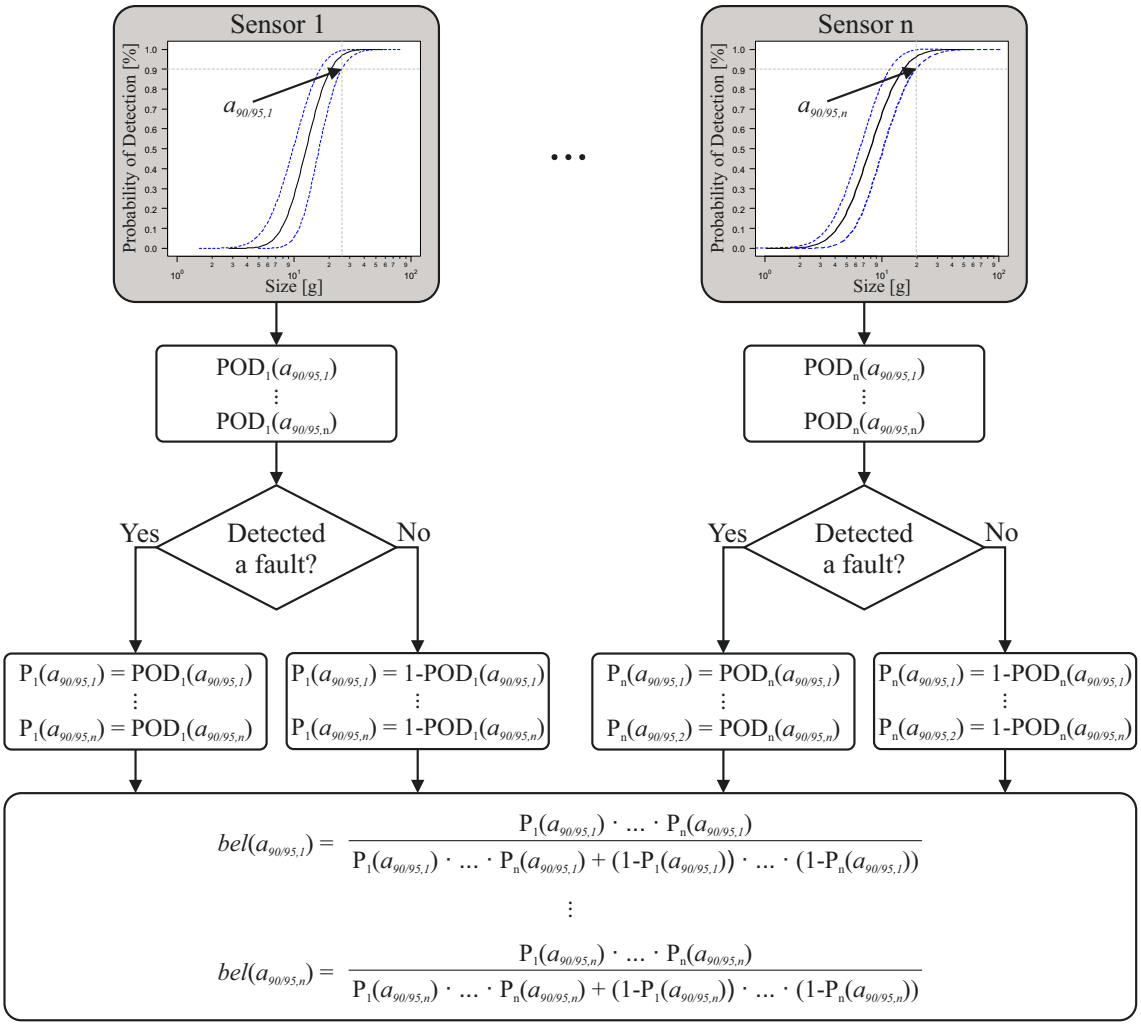

Figure 3.8: Concept of POD-based fusion [Rot19]

3.5 Application to experimental data

The concept for POD-based fusion is applied to the fault diagnosis of an elastic beam. A new approach used to improve the POD characterization of each sensor-/vibration-based statement by decision fusion using several sensors is presented. Here the measurable POD values for specific masses are assumed as replaceable for the precision value. 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. In the case of non reliable statements, the use of one sensor and feature is not suitable to ensure confident health status statements. Decision fusion of the results from different sensors or features may be an option for improving reliability. To fuse the detection results of the sensors related to their POD and $a_{90/95}$ value, the Bayesian Combination Rule (BCR) can be applied. The 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 with $k = 1, ...K$, where K is the total number of considered classifiers is defined as

$$C^k = \begin{bmatrix} C_{11} & C_{12} & \dots & C_{1M} \\ C_{21} & C_{22} & \dots & C_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ C_{M1} & C_{M2} & \dots & C_{MM}, \end{bmatrix}, \tag{3.1}$$

where $i, j = 1, ...M$ with M as the number of classes. The element C_{ij} is the number of samples, where the classifier e_k has assigned class j and the actual class of the sample is i.

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

$$P_{ij} = P(x \in i | e_k(x) = j) = \frac{C_{ij}^k}{\sum_{i=1}^{M} C_{ij}^k}. \tag{3.2}$$

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

$$P^k = \begin{bmatrix} P_{11} & P_{12} & \dots & P_{1M} \\ P_{21} & P_{22} & \dots & P_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ P_{M1} & P_{M2} & \dots & P_{MM} \end{bmatrix}. \tag{3.3}$$

The diagonal values ($i = j$) are the same as the precision value 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

$$bel(i) = \frac{\prod_{k=1}^{K} P_{ij_k}}{\sum_{i=1}^{M} \prod_{k=1}^{K} P_{ij_k}}, \tag{3.4}$$

where j_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 for one of the classes.
In Tab. 3.2 the POD of the seven sensors for the corresponding $a_{90/95}$ values are given. To explain the combination using POD values, the fusion of the sensors ACC 1 and ACC 2 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 values used for belief value claculation are denoted as $P_1 = POD_{20.20g}^{ACC1}$ and $P_2 = POD_{20.20g}^{ACC2}$. The belief value for the mass of 20.20 g is calculated by

$$bel(20.20g) = \frac{P_1 \cdot (1 - P_2)}{P_1 \cdot (1 - P_2) + (1 - P_1) \cdot P_2}. \tag{3.5}$$

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.

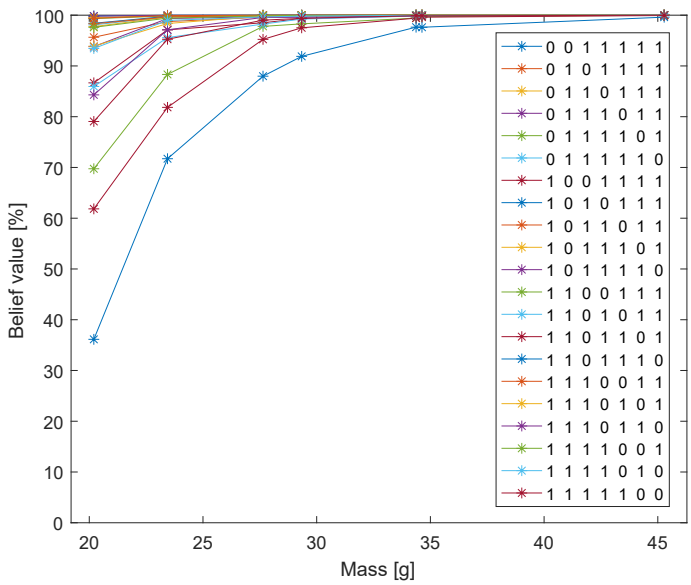

Figure 3.9: Belief values for different detection combinations, when 5 of 7 sensors detect a fault [ARS20]

The resulting belief values for different detection combinations can be established. 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 Fig. 3.9. Depending on which five of the seven sensors detect the fault, the belief values vary. However, all belief values increase for increasing masses. This

Table 3.2: Pod results for case 1 - band power feature [ARS20]

Sensor	POD (20.20 g)	POD (23.44 g)	POD (27.64 g)	POD (29.34 g)	POD (34.37 g)	POD (34.63 g)	POD (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

means that in case of five sensors detecting a fault, the probability that a fault with a higher mass is present is higher . In Fig. 3.10 a selection of different detection combinations is presented. For selection of the best and worst sensor-/feature-based statement, results according to Tab. 3.2 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 Fig. 3.10). For the other cases a symmetry can be seen, e.g. the curve of 5W corresponds to 100 % minus the curve of 2B.
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 sensor-/feature-based statement have detected a fault, it could be, 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.

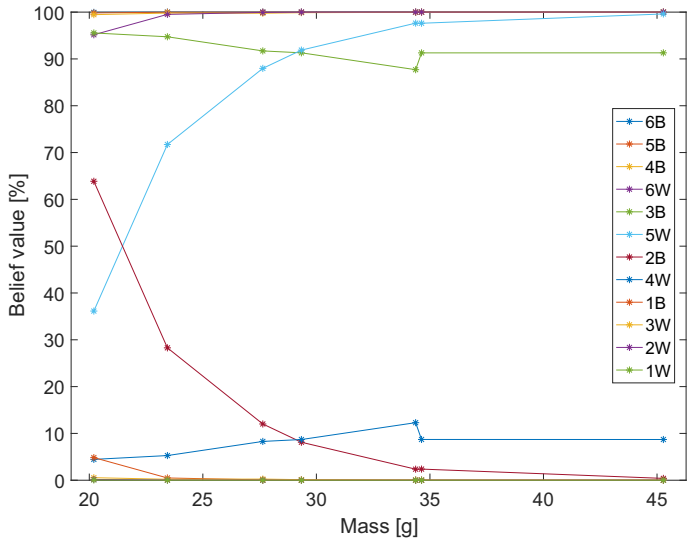

Figure 3.10: Belief values for different detection combinations, selected by the best (B) or worst (W) sensors detecting the fault [ARS20]

Using the introduced fusion approach, the probability of the presence of a specific mass (as fault) can be obtained based on the individual performance and assignments of the sensor-/feature-based statement.

3.6 POD view to FDI

Structural Health Monitoring systems applied to vibrating elastic structures usually denote monitoring dynamical systems. Arising questions are related to the reliability of measurements. Consequently this also affects related diagnostics statements about diagnosis and therefore strongly affects the reliability of vibration-based measurements. The POD-strategy introduced now allows the quantification of vibration-based measurements (here: fault size). Further a new adaption of POD measure is proposed and implemented with respect to the integration of the vibration analysis. Conventional NDT approaches apply statistical analysis and use the relationship between signal strength $\hat{a}$ and the size a of the target initially causing the measurement. The variation of this can be discussed from a FDI-oriented view. In SHM, additional data analysis is required to convey information about the signal's attributes. The response is a feature-based response. The output sensor values for varied fault size have to be recorded. Through signal processing, suitable features have to be extracted and used as response.

The detection task corresponds to the response threshold value ε. This can be obtained from the regression line, confidence bounds, and 90 % probability density

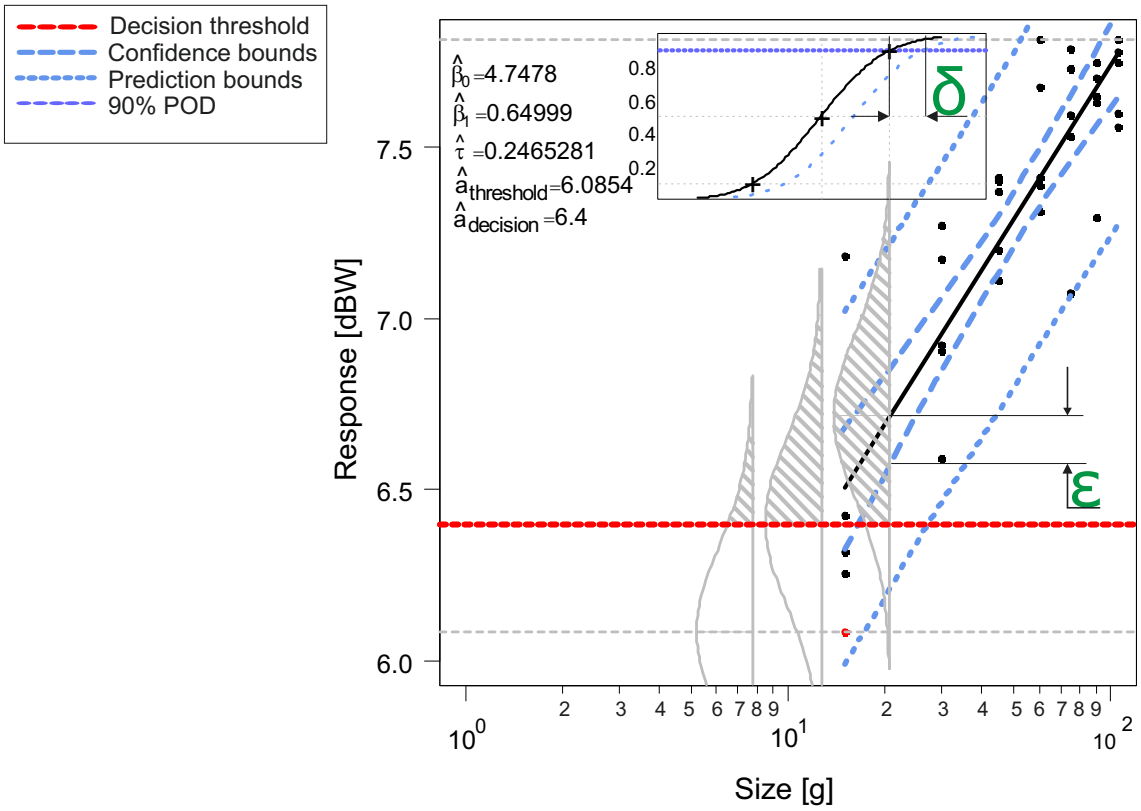

Figure 3.11: Detection task [ARS20]

(Fig. 3.11). From the general equation of the confidence bounds Eqn. (2.2), the values of slope, gradient, and standard deviation can be measured directly from the regression line. However, the size a generating $a_{90/95}$ value can be obtained from the regression line in combination with the POD curve as indicated in Fig. 3.11. The threshold ε defines the response detection value beyond which the fault can be detected with a $a_{90/95}$ reliability. The $a_{90/95}$ value δ quantifies the threshold size. The threshold ε permits reliability certification of the response value for a specific sensor and the subsequent quantification of the response in terms of fault size. The results for this example indicates that a response of 0.14 dBW is the threshold value ε, which corresponds to a fault size δ of 4.88 g. From the demonstrated consideration, it can be stated that this is a new and significant insight for task-/application-specific quantification of sensors and serves as detection and quantification metric for reliability evaluation of sensors. This new view to a classical problem as well as classical solutions should aid monitoring/fault detection system designers to learn about the complexity of the problem and therefore to improve detection systems by choosing the right combination of task, sensor, feature, and sensor position.

3.7 Noise analysis-based discussion

Qualifying sensor diagnostic capabilities for detection and quantification also requires noise analysis. Noise is always present in measured data. Assuming sensor observability, the observed signals aggregates features and characteristics of the monitored system including noise. Noise in this context refers to observed signal with no useful flaw characterization information. In this work a new data driven method is developed where the healthy state data are used to compute the noise.

Classical POD methods usually evaluate independently the noise or infer noise from data not associated with target size [MHA09] [Ann17]. However a new approach is introduced. For illustrative purpose, accelerometer measurements at position P3 filtered as eigenfrequency (mode 1 case II) are used (as shown in Fig. 3.12).

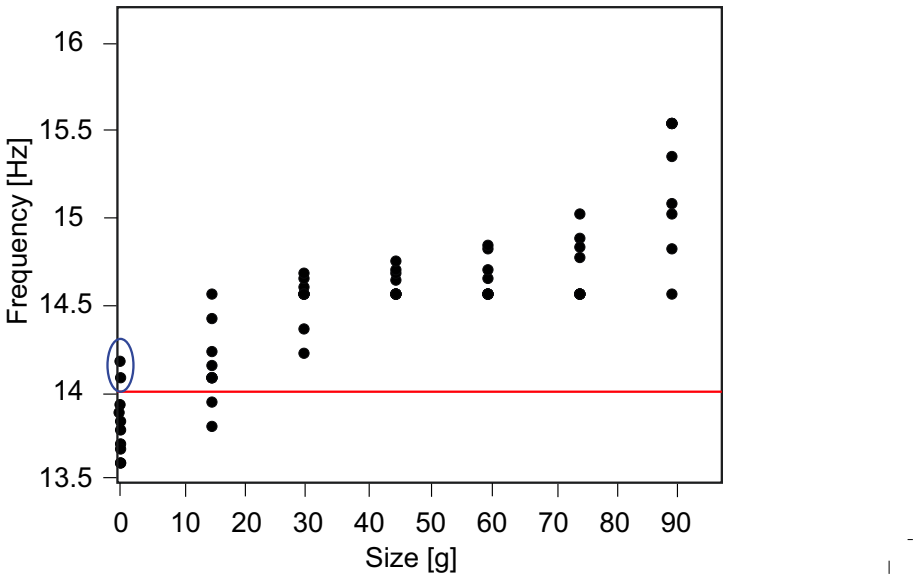

Figure 3.12: Proportion of healthy data above threshold [AS20]

Flaw detection probability and false positive probability are dependent on each other through definition of system decision threshold. The decision threshold signify system response value above which the system is considered faulty. However from Fig. 3.12, it becomes evident that there are still healthy state data (circled points) above the defined threshold to be analyzed. This implies that the region above the threshold cannot be considered to correspond entirely to faulty states. The healthy state data, with no flaw characterization information, can thus be used to determine false positives associated with a selected threshold. The healthy and faulty states data are plotted together and the effect of a selected decision threshold y_{th} on the Probability of False Positive (PFP) for a specific sensor is inferred from the healthy state data. The PFP is the percentage of healthy data that the system wrongly

classifies as damage. The healthy state data of the pristine beam are measured and used as noise.

For noise analysis, first the nature and distribution of the noisy data needs to be established. One technique to ascertain the distribution of noise data is through hypothesis testing.

A χ^2 (chi-squared) test is undertaken to identify the nature of noise distribution. Various distributions are tested with the Gaussian distribution emerging most plausible. The χ^2 produced a p-value of 0.95 thereby rejecting the null-hypothesis that the distribution is non-Gaussian. A regression analysis is carried out on the noisy data and the mean μ_{noise} and standard deviation σ_{noise} are calculated (Fig. 3.13). For a Gaussian noise distribution, the PFP is computed as

$$PFP = \int_{y_{th}}^{\infty} \frac{1}{\sqrt{2\pi}\hat{\sigma}_{noise}} e^{-\frac{(y-\hat{\mu}_{noise})^2}{2\hat{\sigma}^2_{noise}}} dy.$$

The distribution with regards to PFP is illustrated in Fig. 3.14 (shaded red area relative to the selected decision threshold).

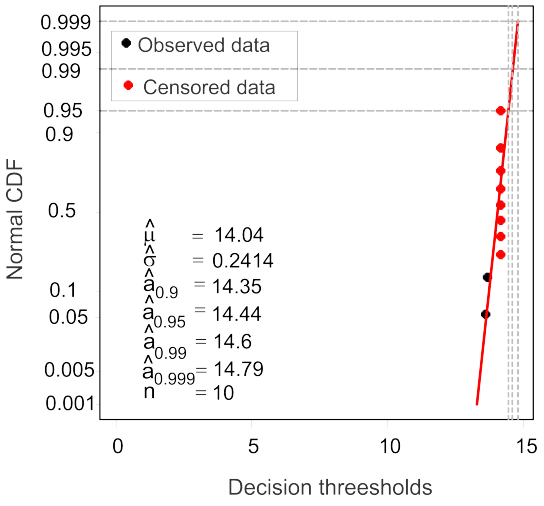

Figure 3.13: Mode 1 accelerometer at P3 noise data [ARS20]

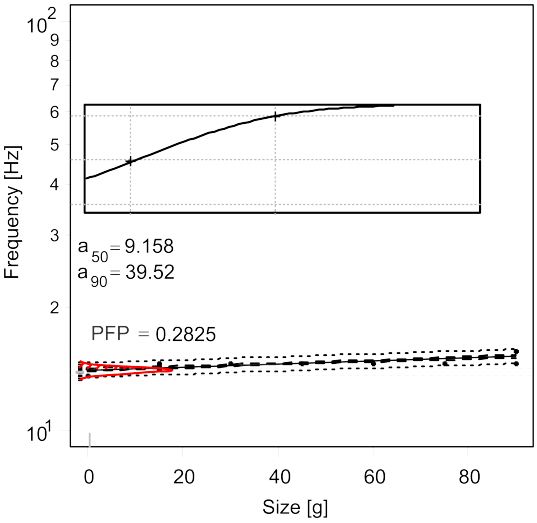

Figure 3.14: POD and corresponding PFP [ARS20]

The distribution regarding PFP and the corresponding POD values for accelerometer measurements at position P3 is illustrated in Fig. 3.13.

For every selected cutoff point a single PFP value is evaluated but detected flaw size is dependent on the probability level (50 %, 90 %, etc.). In this work the 90 % and 90/95 POD sizes are used (similar to typical standards in industry [MHA09]). Changing the sensor results in changing the distribution character. Other sensors

Table 3.3: Tradeoff between decision threshold, PFP, and 90 % POD [AS20]

Sensor	13.2	13.4	13.6	13.8	14.0	14.2	14.4	14.6	14.8
	1.000	0.996	0.966	0.841	0.568	0.255	0.0685	0.0103	0.0008
ACC 1	0.00	0.00	0.00	5.85	23.36	40.62	57.64	74.43	90.99
ACC 2	0.00	0.00	0.00	7.64	24.53	41.18	57.59	73.78	89.75
ACC 3	0.00	0.00	0.00	7.64	24.53	41.18	57.59	73.78	89.75
SG 1	0.00	0.00	0.00	7.64	24.53	41.18	57.59	73.78	89.75
SG 2	0.00	0.00	0.00	6.71	24.38	41.80	58.98	75.92	92.63
Las 1	0.00	0.00	0.00	3.71	21.90	39.83	57.52	74.96	92.16
Las 2	0.00	0.00	0.00	7.64	24.53	41.18	57.59	73.78	89.75

Table 3.4: Tradeoff between decision threshold, PFP, and 90 % POD [AS20]

Sensor	13.2	13.4	13.6	13.8	14.0	14.2	14.4	14.6	14.8
	1.000	0.996	0.966	0.841	0.568	0.255	0.0685	0.0103	0.0008
ACC 1	0.00	0.00	0.418	15.83	31.46	47.68	64.79	82.74	101.13
ACC 2	0.00	0.00	1.79	16.82	32.06	47.80	64.32	81.55	99.15
ACC 3	0.00	0.00	1.79	16.82	32.06	47.80	64.32	81.55	99.15
SG 1	0.00	0.00	1.79	16.82	32.06	47.80	64.32	81.55	99.15
SG 2	0.00	0.00	2.10	17.40	33.1	49.40	66.80	85.10	103.9
Las 1	0.00	0.00	0.00	14.54	30.61	47.32	65.08	83.80	103.05
Las 2	0.00	0.00	1.79	16.82	32.06	47.80	64.32	81.55	99.15

will result in different distribution characters and different values; however, the character of the introduced strategy remains. The procedure is repeated for all sensors for the same feature (here: mode 1 eigenfrequency case II). A summary of the 90 % and 90/95 POD and PFP values for a selected decision threshold are illustrated in Tab. 3.3 and 3.4 respectively. The results from Tab. 3.3 and 3.4 shows the characteristic of the flaw (here: size) affects the detection probability. This important but often ignored outcome (especially in ROC and PR evaluation), shows the need for a parametric evaluation system incorporating the effect of flaw characteristic in the evaluation process.

3.7.1 Improved POD analysis

A new method to improve the detection capabilities of sensors is introduced in this section. The main idea is to discuss the sensitivity of features to the fault to be detected. The POD results comparing different sensor/feature combinations indicate that the most sensitive feature to point mass changes is band power. Additionally, by utilizing the introduced noise analysis procedure, direct comparison can be made between the eigenfrequency and band power for the same sensor (here: accelerometer

at position P3). The noise data and corresponding PFP for band power feature for accelerometer 3 are shown in Fig. 3.15 and 3.16.

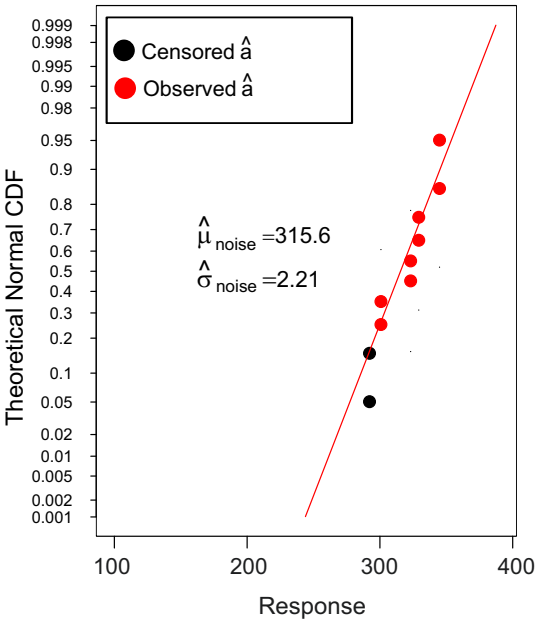

Figure 3.15: Noise data of band power feature of accelerometer at P3 [AS20]

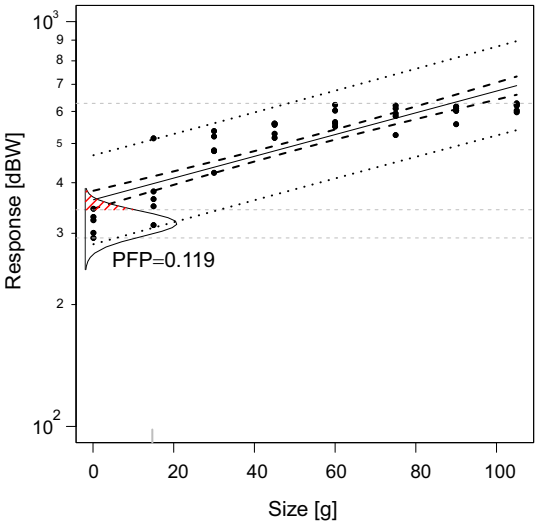

Figure 3.16: POD and corresponding PFP for accelerometer 3 [AS20]

The trade-off between PFP, 90/95 POD, and flaw size can then be constructed for both features (eigenfrequency mode 1 and band power) as illustrated in Fig. 3.17.

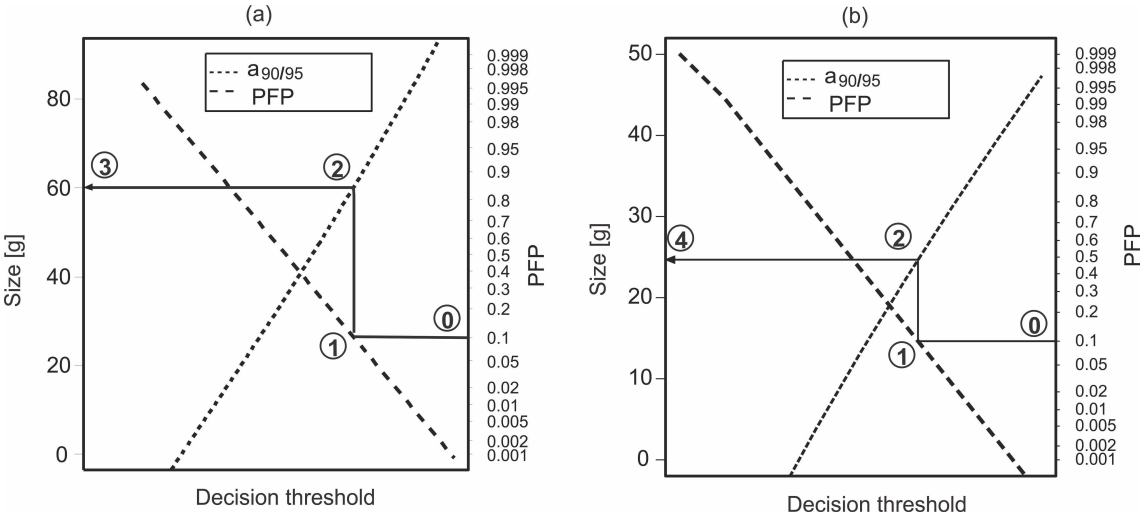

Figure 3.17: Tradeoff between PFP, POD and flaw size (a): Eigenfrequency (b): Band power [ARS20]

The analysis indicates for a selected PFP value of 0.1 (1) at 90 % probability and 95 % reliability level (2), the eigenfrequency feature detects a flaw of 60.2 g (3) while the band power feature detects a flaw size of 24.9 g (4). This procedure provides an effective and simple method to improve the detection capabilities of sensors without affecting (here increasing) the false alarm rate.

3.8 Summary

This chapter focuses on introducing a novel POD-oriented view to vibration-based diagnosis typically using sensor types. The measurements used are acceleration, strain, and displacement (laser sensors). Challenges associated with monitoring dynamical systems are presented. The results indicate that the POD characterization depends on the sensor position, fault position, and the feature selected. The sensor type has an effect on the POD due to the fact that performance specifications vary for different sensors. The $a_{90/95}$ criteria representing probability of 90 % at a confidence level of 95 % is successfully implemented in vibration-based FDI as a reliability measure. The new insight introduced allows task-/application-specific quantification of sensors relative to vibration-based monitoring/diagnosis of faults. Using the new 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. Noise analysis allows a decision threshold to be selected which permit a suitable trade-off between the POD and PFP. The flaw characteristic (here:size) plays a crucial role in its detection and the introduced approach provides a means to assess that.

4 New evaluation assessment of Machine Learning classification approaches

This chapter presents a new evaluation assessment of machine learning approaches. A ML approach may be suited for a specific task depending on the application and data set. To select an approach for a task, performance evaluation may be imperative. Typical evaluation approaches utilize the receiver operating characteristic and precision-recall curves. Despite the popularity and acceptance of these two classification evaluation approaches, they do not address the influence of process parameters on classification results. Both curves are unable to quantitatively relate detectability to a process parameter. In this chapter this limitation is discussed and addressed by adapting the POD reliability measure but with a different implementation technique. The chapter consist of three parts. Firstly a new visualization of the effect of process parameter on the classification results is illustrated. Consequently it can be established that any selected decision threshold produces a single false alarm rate value, however the detection rate values varies contingent on the process parameter. Secondly, comparison of different classifiers utilizing the new POD approach is presented. This is useful because the POD method permits a parametric comparison of classifiers and hence expedient when interpreting complex sensor data describing a complex spatial scenario. Finally, improvement of the classifiers detection capabilities through the use of a test-specific target response measure is presented. Based on the newly introduced POD-related evaluation, different classifiers can be clearly distinguished with respect to their ability to predict the correct driver behavior as a function of remaining time before the event itself.

The content, figures, and tables in this chapter are based on the publication [ADS19] and the submitted papers [ADS20a] [ADS20b].

4.1 Introduction

Machine learning has been developed and used for decades in many applications. Advancements in processing capabilities of computers has made ML implementation relatively faster. Though used mainly in science and technology, it has been embraced by the humanities. Institutions and researchers are using ML to analyze data patterns to predict consumer behavior, detect frauds in financial transactions, clinical trials, and statistical analysis to literary works [LAGS+13]. Several ML approaches exist. An approach may be better suited depending on the data set and application. To select an approach, performance evaluation may be useful. The ROC and PR curves are among the commonly used evaluation tools. Both curves provide graphically standard tools to evaluate the performance of a binary classifier as its discrimination threshold is varied. Whilst the ROC curve uses the ratio of DR

to FAR, the PR curve utilize the ratio of precision to recall. These evaluation curves can be used to determine optimal decision thresholds and allow graphical comparison of diagnostic tests. The final evaluation process of both curves provide a measure to select optimal models discarding properties related to process parameters on classification results. This implies, these curves cannot quantitatively relate detectability to a process parameter [Ann17] [MHA09]. Secondly the ROC curve does have an additional limitation of not considering the prevalence [MHA09]. The prevalence considers the total number of cases of targets in a given sample/population at a specific time. Conventional justification for the ROC is based on the overlapping probabilities for signal and noise from measurements (see Fig. 4.1). Therefore, ROC advocates argue that a trade-off between DR and FAR describes fully the test results.

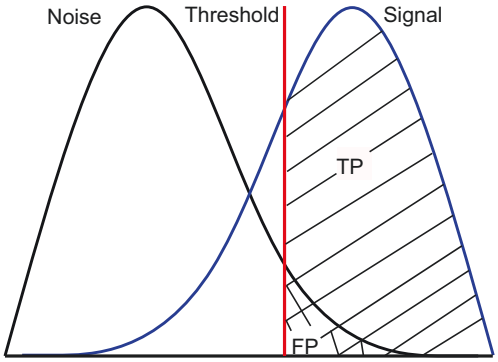

Figure 4.1: Overlapping Probability densities of signal and noise

Many authors have critique the use of area under ROC (AUC) as it tends to be dominated by the high FPR points [BKD01] [LAGS$^+$13]. Though the AUC is buoyed by observations to the right, highly-ranked transactions occur on the left side (region with the least AUC) of ROC [LAGS$^+$13]. Difficulties with the ROC when dealing with highly skewed datasets is reported in [LAGS$^+$13]. Some authors propose PR curves as more informative than ROC when dealing with highly skewed datasets. Research has mainly focused on addressing skewed datasets and using portions of AUC [LAGS$^+$13]. However, the fundamental concern of investigating the effect of process parameter on classification results has received little attention. Despite these shortcomings the ROC and PR are still widely used largely because few alternatives exist [MHA09]. This chapter attempts to address this concern using the Probability of detection reliability measure. The POD will be adapted, modified, and implemented in assessing classifier performance by relating detectability quantitatively to a process parameter in the evaluation.
In this chapter the POD approach is used to evaluate classifier performance incorporating time t as a process parameter. The proposed approach is used to compare the suitability of Artificial Neural Networks (ANN), Hidden Markov Models (HMM), Random Forest (RF), Support Vector Machines (SVM), and modified versions of

these classifiers with respect to the reliability of related classification of upcoming events prior to the real ones, here: human decisions based on the visual impression of changes within the environment. Consequently, based on the POD-related evaluation the different classifiers can be clearly distinguished with respect to their ability to predict the correct behavior as a function of time prior to the event itself. This allows a very detailed comparison of classifiers related to the effects of parameters affecting the processes to be classified.

4.2 LANE CHANGE PREDICTION USING CLASSIFIERS

A very brief presentation on the ML algorithms used to predict driver intention recognition is detailed in this section. The authors in [DS19] proposed a strategy to improve training of conventional algorithms. The authors showed that usually a set of unknown classifier tuning parameters are needed to be set manually before training, when a conventional algorithm is used. With the proposed training procedure, the most suitable values of these unknown parameters can be determined automatically to optimize the performance of the conventional algorithms. The authors compared the reliability of the algorithms with respect to the relevant accuracy, detection rate, and false alarm parameters. In this contribution, eight classifiers; conventional/modified SVM, HMM, ANN, and RF are used. These classifiers prediction abilities will later be evaluated using the POD method.

4.2.1 Conventional/modified Support Vector Machine

Support Vector Machine is a supervised machine learning method and is a widely applied classification technique [CV95]. The various classes in SVM refer to different driving behaviors, which are classified by transforming the observation variables into an observation vector and thus generating a distribution in a high dimensional space. Each observation vector is assigned to corresponding classes based on training data. The three driving behaviors to be predicted are separated using hyperplanes. The process of SVM learning is trying to find an optimal hyperplane between observation vectors of different classes to generate a maximal geometric margin. However, the SVM was originally designed only for two classes. In this contribution, the driver prediction model is a multiclass problem (lane change to right S_1, lane keeping S_2, and lane change to left S_3). To solve this complexity one-against-one approach is utilized to establish a multiclass model. Data processing is not needed for the conventional SVM, it can be trained with raw data. An improved SVM is trained with a prefilter [DS19] [DS18] applied to define features and influence the prediction performance of SVM.

4.2.2 Conventional/modified Hidden Markov Models

An HMM [Rab89] describes the relationship between two stochastic processes: one consists of a set of unobserved (hidden) states which cannot be measured directly. The other stochastic process is denoted by observable symbols. Due to dynamical changes, the process considered moves from one state to another generating hidden states sequence and observation sequence. In a given observation sequence and its corresponding hidden state sequence, the HMM parameters can be calculated (denoted as training) and adjusted to best fit both sequences using Baum-Welch algorithm. Based on the obtained HMM, the most probable sequence of hidden states, which has the highest probability, can be calculated by using Viterbi algorithm. In this work, three different driving behaviors (S_1, S_2, S_3) are modelled as hidden states for the HMM. The related training process method of conventional HMM is referred to [DWS18] and an improved HMM is trained with an optimal prefilter [DS19].

4.2.3 Conventional/modified Artificial Neural Network

Artificial Neural Network is a computational model that imitates biological neural network. These models learn to perform tasks without explicit task-specific rules. Typically, ANN contains many layers, the first and last layers are the input and output respectively with signals traveling from input to output. To determine final results, usually cut-off thresholds are used to distinguish the decimal values into class labels. The values of cut-off thresholds are unknown and should be set prior to the training. Related parameters of a conventional/improved ANN [DS19] are used.

4.2.4 Conventional/modified Random Forest

Random Forest is an extension of decision tree method and it is used to solve classification or regression problems [Bre01]. The RF algorithm contains a set of randomized decision trees that are independent from each other. Each decision tree is trained by a randomly selected Bootstrap sample [Bre01], which is generated from the training data set with replacement. The total number of decision trees N_{Tree} should be defined before the training process. However, the number of N_{Tree} is unknown. In [OPB12], the authors pointed out that the value of N_{Tree} is worth optimizing. Therefore, an improved RF is trained with optimal parameters (including prefilter thresholds and N_{Tree}) defined in [DS19]. Similarly, raw data and default N_{Tree} are used to train a conventional RF.

4.2.5 Classifier evaluation using POD

The proposed POD model building process to evaluate ML approaches performance is illustrated in Fig. 4.2. The procedure involves:

1 Identify process parameter by which detectability can be related to quantitatively

2 Identify ML approaches/classifiers to be used for classification

3 Determine if a conventional or optimized classifier is used (procedure used in this work illustrated)

4 Train classifiers

5 Generate prediction model

6 Compute performance measure

7 Implement POD approach (Target response or Hit/Miss) based on nature of data

8 Evaluate POD

The POD is thus evaluated and generated based on the above mentioned procedures from step 1 to 8.

4.3 EXPERIMENTAL RESULTS

The experimental setup and data of the work of Deng [DWS18] are used in this section, however the research methods and applications presented are completely new. As example, the comparison of ML approaches applied to driver intention prediction is used. Here the reliability of outcomes of classifiers is examined by using the POD approach. Therefore the POD measure has to be adapted and implemented. Using data from real human driver simulator the proposed approach is demonstrated and applied to different ML classification approaches. Additionally, the classifiers are compared using the proposed approach. The obtained results and the analysis are discussed in detail.

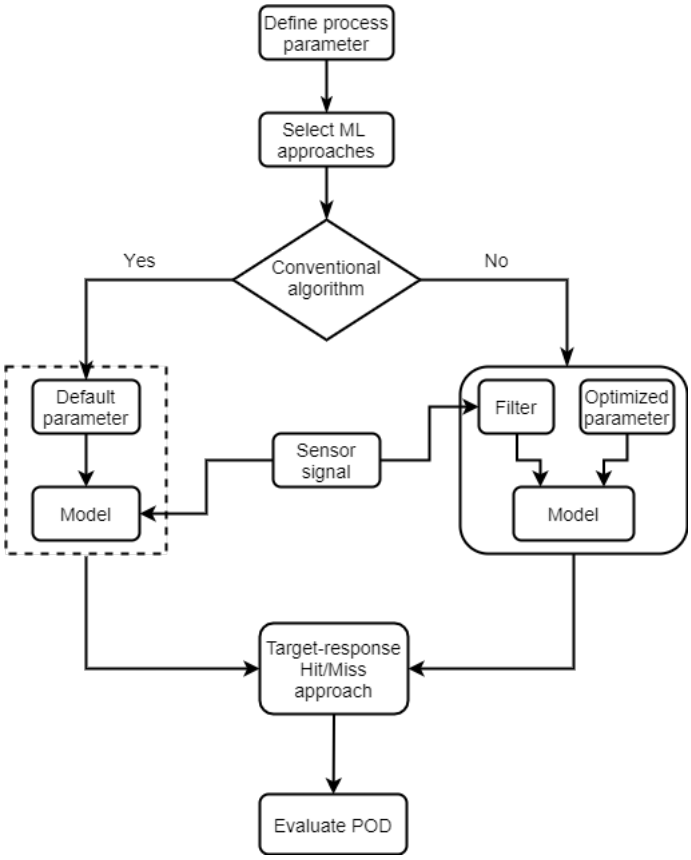

Figure 4.2: Classifier POD model building process [ADS20a]

Figure 4.3: SCANeR^{TM} studio, Chair of Dynamics and Control, UDuE, Germany [ADS20b]

4.3.1 Experimental set-up

The driving simulation is performed using $SCANeR^{TM}$ studio driving simulator (Fig. 4.3). The simulator is equipped with five monitors, base-fixed driver seat, steering wheel, and pedals. The three rear mirrors are essential to decide lane

change and are displayed on the corresponding positions of the monitors.

A typical driving scenario is illustrated in Fig. 4.4 and requires the ego driver (red vehicle) to make a decision. Correct decisions are made with the aid of several sensor data. The driving simulator simulation engine and corresponding input sensors

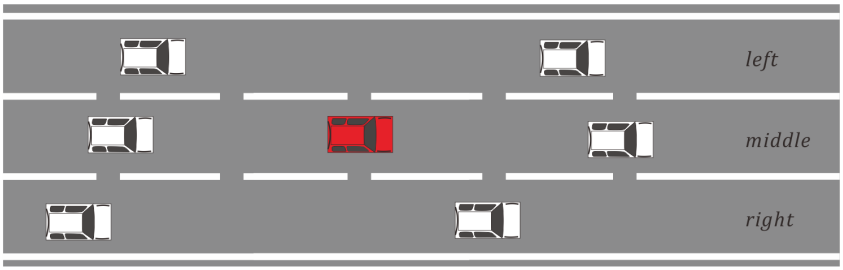

Figure 4.4: Driving environment [ADS20b]

(Fig. 4.5) aids in decision making. For example, vehicles can use camera images to find lane lines or track other vehicles on the road. Like how humans perceive the world, vision-based data provided by the simulator aids vehicles to recognize traffic lights, traffic signs, and speed signs. With all the information collected, driving assistant system (human behavior prediction model) can be established, and finally suggestions/warnings are given to driver to control the vehicle's direction, speed among others.

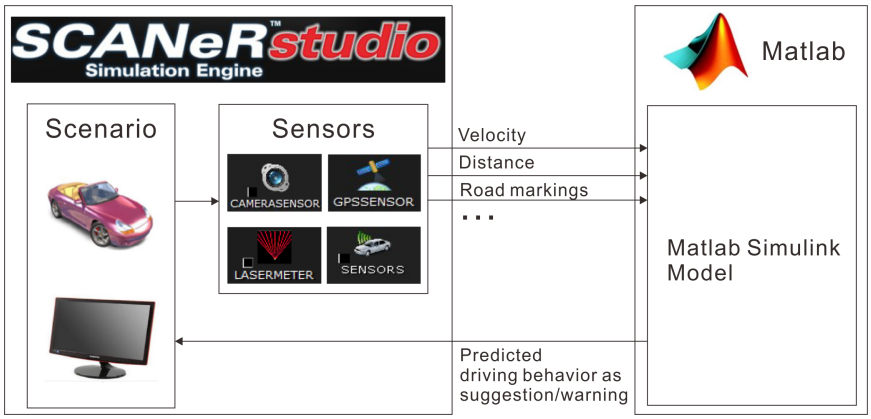

Figure 4.5: SCANeR^{TM} studio simulation engine [ADS20b]

The driving environment is a highway-based traffic scenario with four lanes of two directions and simulated traffic environment. During driving, the participant could perform overtaking maneuver when the preceding vehicle drives slowly. After overtaking, the participant is permitted to stay on the new lane or drive back to the initial lane. The time points of changing lane to left and right are decided by the participants. The lane change/keeping behaviors are illustrated in Fig. 4.6.

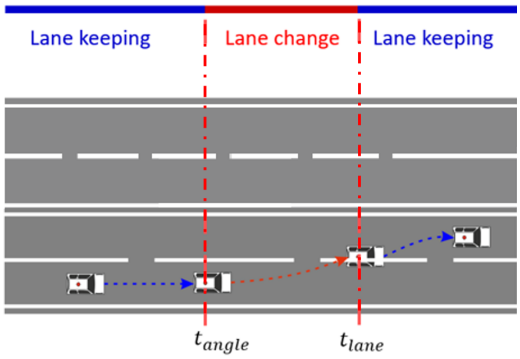

Figure 4.6: Lane changing behavior [ADS20b]

4.3.2 Data processing

Eight classifiers are selected to train driving behaviors prediction models in this article. As detailed in section 4.2, the algorithms are used to train conventional and modified models. The driving behaviors prediction model based on the classifiers is shown in Fig. 4.7. It consists of two processes: parameter definition and driving behaviors prediction.

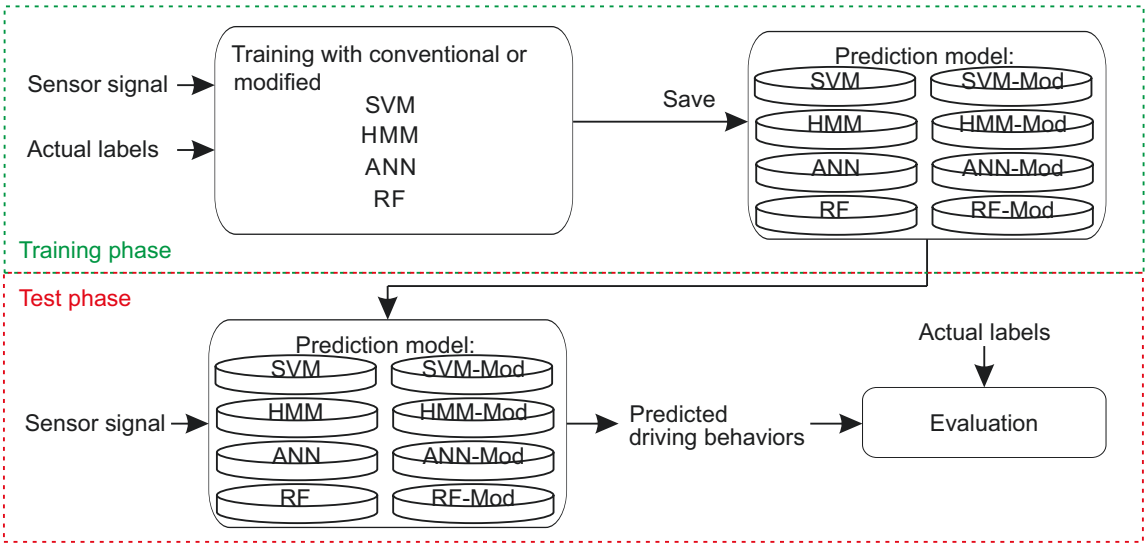

Figure 4.7: Test/training model [ADS20b]

4.3.3 Target response value

The target response approach is adapted and implemented in this work because the data to be analyzed relates a changing parameter to a response value quantitatively. The approach is appropriate in this context due to a relationship between

a changing parameter (here:time) and a changing response function. To generate the response value, the evaluation of the classifier relative to a performance is calculated. Conventional and modified models are calculated in the training phase. Based on these models, the driving behaviors in the upcoming driving processes can be determined. The measured and estimated driving behavior states are compared to check the correspondence. To evaluate the performance of classifiers, accuracy is one of the most commonly used metric due to its simplicity. However accuracy does not precisely describe the classifier performance when dealing with unbalanced data. A high accuracy can be achieved by correctly classifying the majority class while neglecting the minority class. To avoid this, in this contribution F-measure/score (FM) is selected as target response value which is calculated using the multiclass confusion matrix (Fig. 4.8).

Multiclass Confusion Matrix		Predicted		
		S_1	S_2	S_3
Actual	S_1	TP	FN	FN
	S_2	FP	TN	TN
	S_3	FP	TN	TN

Figure 4.8: S_1 multiclass confusion matrix [ADS19]

The FM value is the harmonic mean between detection rate and precision (also called predictive positive value (PPV)). The DR value is calculated based on TP value and the FN numbers. The PPV is calculated based on TP and FP values.

The DR, FAR, PPV, and FM values are defined by [MMBC13]

$$DR = \frac{TP}{TP + FN}, \tag{4.1}$$

$$FAR = \frac{FP}{TN + FP}, \tag{4.2}$$

$$PPV = \frac{TP}{TP + FP}, \tag{4.3}$$

and

$$FM = \frac{2 * DR * PPV}{DR + PPV}. \tag{4.4}$$

To evaluate the predicted lane change performance, initial ideas of the authors [ADS19] are used. Here, each lane change behavior is defined as a separate event. From 7 seconds before to the time of actual lane change is considered. The time interval is divided into 140 time points, i.e. every 0.05 s. These time points are defined as "recognition time points", and for each time step, FM value will be calculated. The computed FM value at each time point is used as the response value in the POD evaluation.

4.3.4 POD generation process

Based on the computed F score response values at each recognition time point, the target response method is utilized in this section to generate POD for the classification results. The aim is to establish a POD characterization to illustrate the effect of process parameters (here: time) to the classification/fused results so the probability distribution is derived. Many authors intuitively use $log\ x$ vs. $log\ y$ scale in generating the POD but that is based on convention rather than effective statistical modeling [Ann17].

Four models comprising combinations of logarithmic and linear scales (Fig. 4.9) are established for each ML algorithm to ascertain model having [KNN+05] [MHA09]

1 Linearity of the parameters: $E(y_i|X) = x_i\beta$, where x_i is the $i-th$ row of X,

2 Uniform variance: $var(y_i|X) = \sigma^2, i = 1, 2, 3, ..., n$ and

3 Uncorrelated observations: $cov(y_i, y_j|X) = 0, (i \neq j)$.

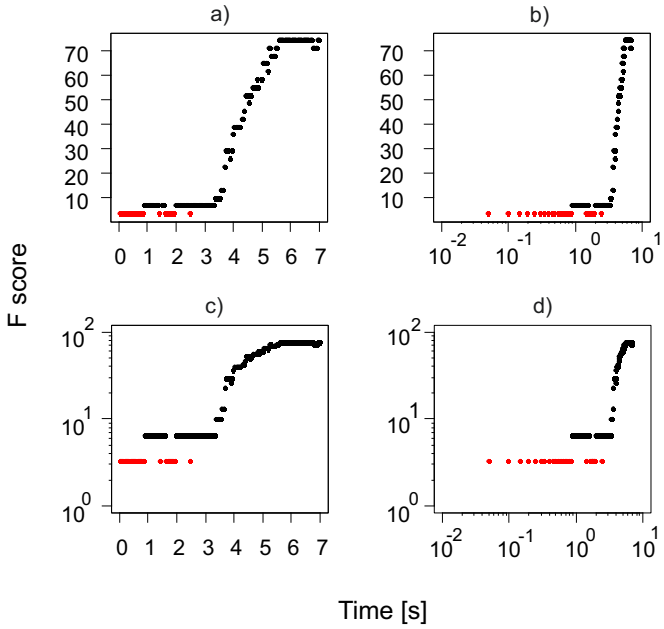

Figure 4.9: Regression models a: x vs. y b: log x vs. y c: x vs. log y d: log x. vs. log y [ADS20a]

In this concrete example the model satisfying the above criteria best is the graph with logarithmic abscissa and linear ordinate (model b) and hence selected for further analysis.

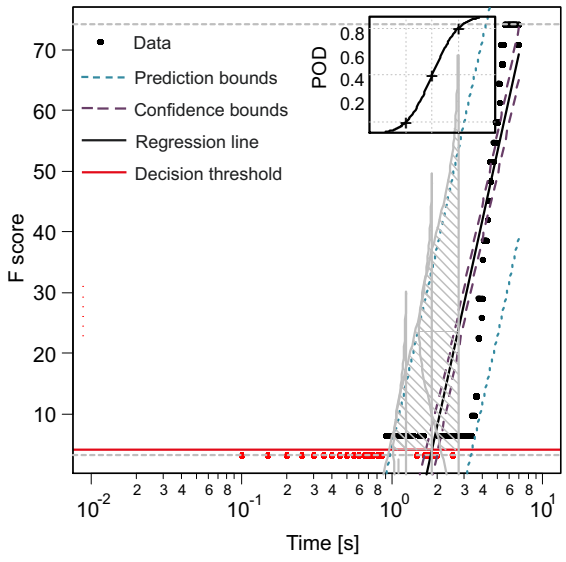

Figure 4.10: Relationship between decision threshold and POD

Regression analysis is implemented on the selected model and the inspection threshold (minimum detectable signal), saturation threshold (maximum inspection thresh-

old for signal), decision threshold (response value above which the signal is considered in the evaluation process), confidence bounds, and prediction bounds are constructed using the formulation from section 2 as illustrated in Fig. 4.10.

The cumulative density function for the data distribution are also constructed. The POD curve is generated using area of the cumulative density function above decision threshold.

The POD curve is analogous to the regression line . The confidence bounds about the regression line are used in the same way to construct the 95 % bounds around the POD curve. For this work, estimation is made for both left and right lane change detection. All the eight classifiers are used in the estimation process. The left and right lane HMM POD curves are shown in Fig. 4.11.

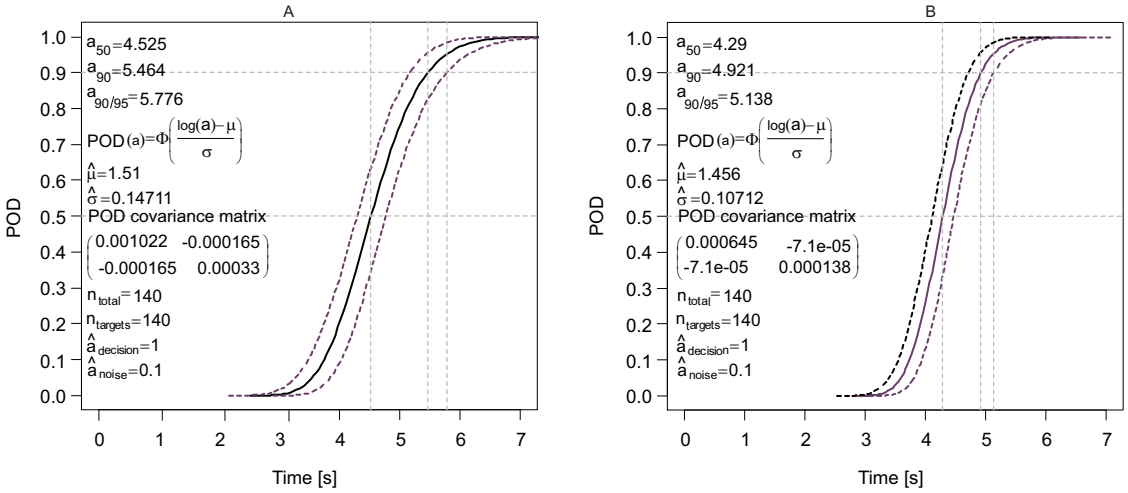

Figure 4.11: POD curves A: left lane HMM curve B: right lane HMM curve

4.3.5 Noise analysis

The observed data aggregate the characteristics of the targets signature corrupted by aberrant signals generally referred to as noise. Classical POD methods measure noise as part of planned experimental measurement, however that is absent in the current work. Noise therefore will be inferred from the observed data. Noise in this context refers to observed signals with no useful target characterization information. Therefore observed data outside the prediction bounds will be interpreted as noise. Still using data from Fig. 4.10, the extracted noise is shown in Fig. 4.12. Statistical χ^2 (Chi-squared) hypothesis test is undertaken to identify the nature of noise distribution.

Various distributions are tested. The Lognormal distribution emerging most plausible. Analysis is carried out on the noisy data and the mean μ_{noise} and standard

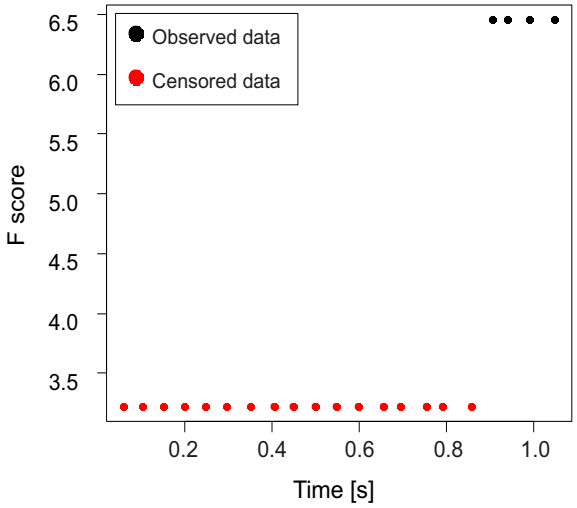

Figure 4.12: Noise data

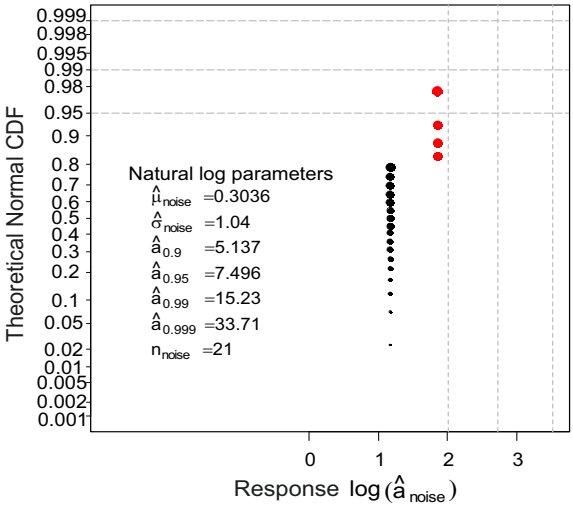

Figure 4.13: Noise paramters

deviation σ_{noise} are calculated (Fig. 4.13). For a Lognormal noise distribution, the FAR or PFP is computed as

$$FAR = \int_{y_{th}}^{\infty} \frac{1}{y\hat{\sigma}_{noise}\sqrt{2\pi}} e^{-\frac{(\ln y - \hat{\mu}_{noise})^2}{2\hat{\sigma}^2_{noise}}} dy.$$

The distribution with regards to FAR is illustrated in Fig. 4.13 (shaded red area relative to the selected decision threshold).

From Fig. 4.14 it becomes evident that for a selected decision threshold, a corresponding unique FAR value exist however the detection probability varies relative to a parameter (here: time). This implies, for a selected cut-off point there is not one FAR to one DR value but one FAR to many DR values. This consideration was not factored initially in the ROC construction because the size of opposition objects during WWII was irrelevant, but modern applications are concerned with how the characteristics of target change the probability of detecting it. To visualize the changing probability distribution with changing thresholds, other cut-off points are selected and the nature of the distribution illustrated in Fig. 4.15. The FAR values for the selected decision thresholds are shown in Tab. 4.1.

Table 4.1: FAR results for different decision thresholds [ADS20a]

DT [%]	4	8	12	16
FAR	0.1489	0.0439	0.0180	0.0088

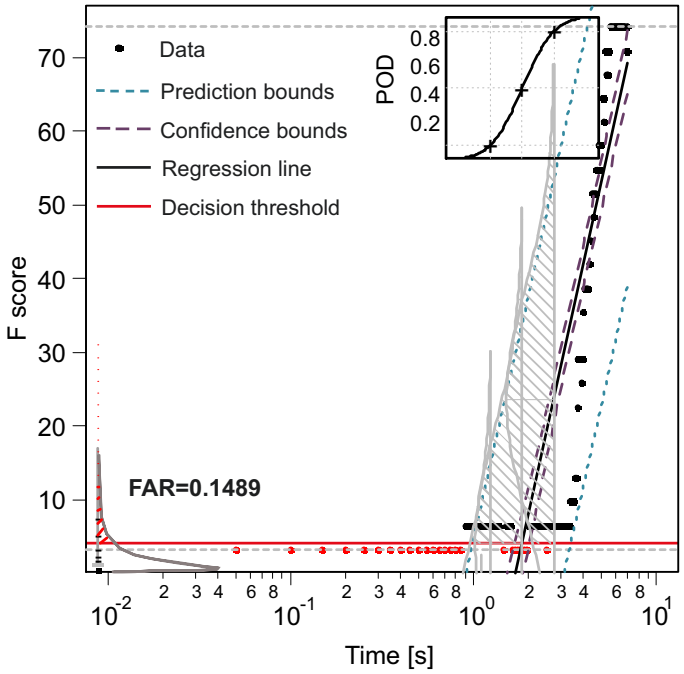

Figure 4.14: Probability of false positive value

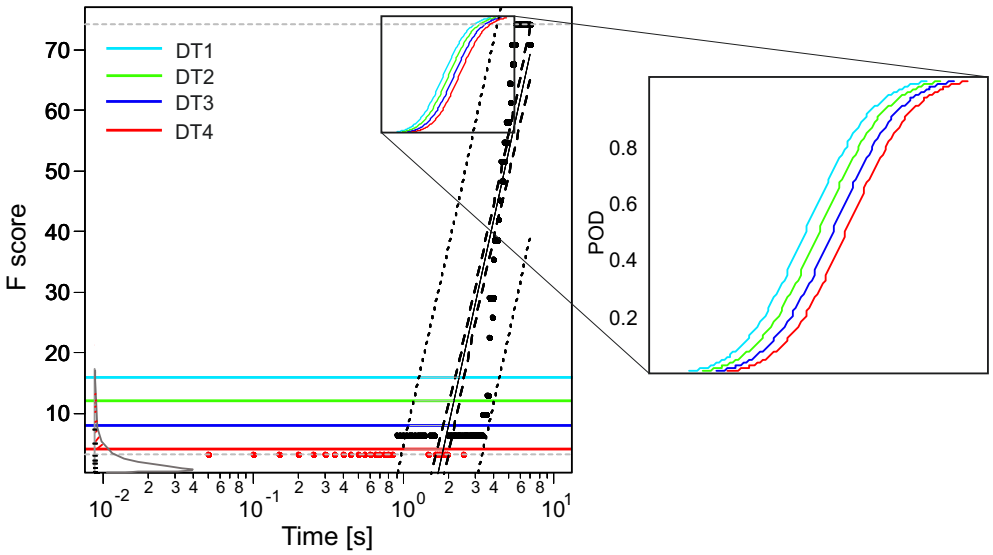

Figure 4.15: Relationship between decision thresholds and POD [ADS20a]

4.3.6 Novel evaluation measure integrating process paramater

In this section, for the first time a new evaluation method concurrently considering the decision threshold, FAR, POD, and process parameter is developed. To illustrate the new approach the generated POD curves for the selected decision threshold in section 4.3.5 are used.

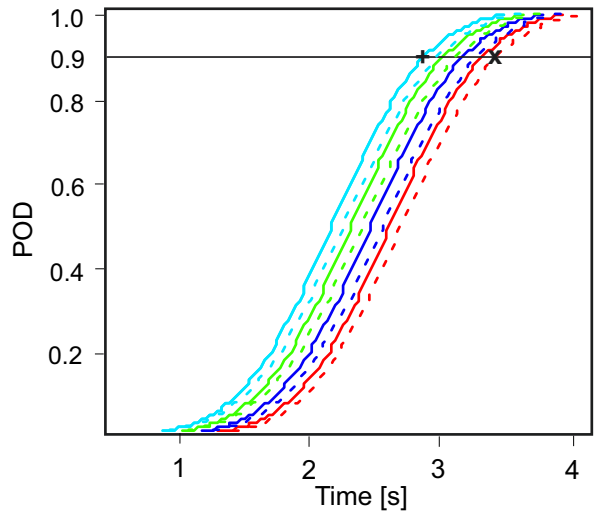

Figure 4.16: Detection at 90 % and 90/95 level, +: 90 % POD **x**: 90/95 POD [ADS20a]

To demonstrate the novel technique, the 0.9 probability is drawn to intercept POD and confidence curves at the points (+) and (**x**) respectively as shown in Fig. 4.16. At the point of intersection; the POD, FAR, decision threshold, and process parameter (here: time) are known. These values are considered for every point on the drawn 0.9 probability line. A graph depicting the relationship between the POD, FAR, DT, and time is shown in Fig. 4.17.

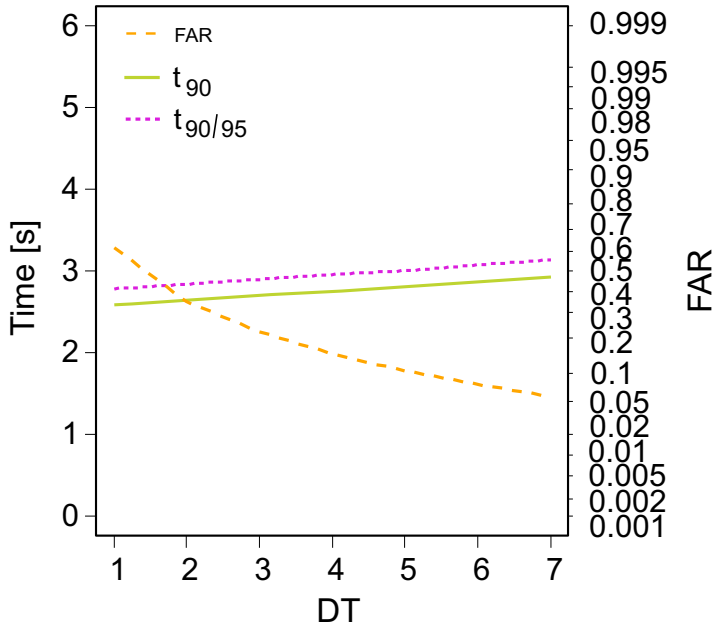

Figure 4.17: Evaluation measure incorporating process parameter [ADS20a]

This is very useful because it provides a measure to incorporate and assess the effect of process parameters on the evaluation process. The Fig. 4.17 correspond to only the 0.9 probability. To see the relation for every probability point requires the procedure to be evaluated for the entire probability range; from 0 to 1.0. The graph corresponding to the entire probability range utilizing the developed method is indicated in Fig. 4.18

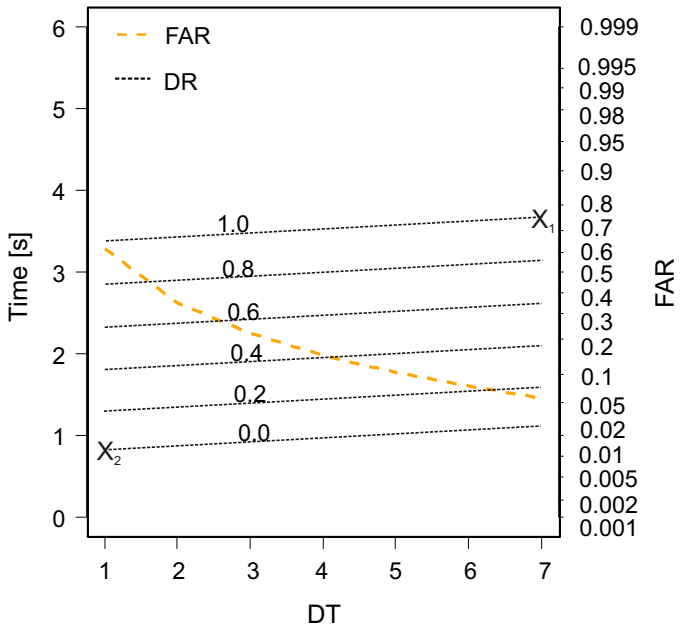

Figure 4.18: Parametric evaluation measure [ADS20a]

The entire POD range contains detection probabilities from 0.0 DR up to 1.0 DR. From the illustrated relations two points can be easily taken. Here X_1 denotes the optimal point corresponding to least FAR and maximum DR. The point X_2 denotes the worst point corresponding to maximum FAR and least DR. However there is an associated cost whereby the point X_1 has a high threshold of 7 % and predicts the impending event at 3.82 s. The point X_2 on the other hand has least threshold value of 1 % and predicts the impending event at 0.8 s. The introduced method presents a novel and significant approach to concurrently examine the detection probabilities, the false alarm rate, the decision threshold, and the process parameter.

4.4 Central outcome: comparison of classifiers using POD

Based on the developed POD approach, eight different classifiers ability to predict driver lane change behavior are examined. The predicted driving behavior as a function of target response at each recognition point for the same selected threshold are shown in Fig. 4.19 and Fig. 4.20.

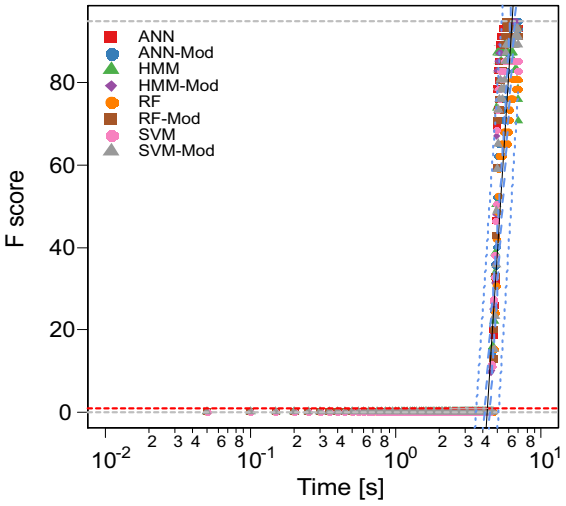

Figure 4.19: Left lane change classifier data [ADS20a]

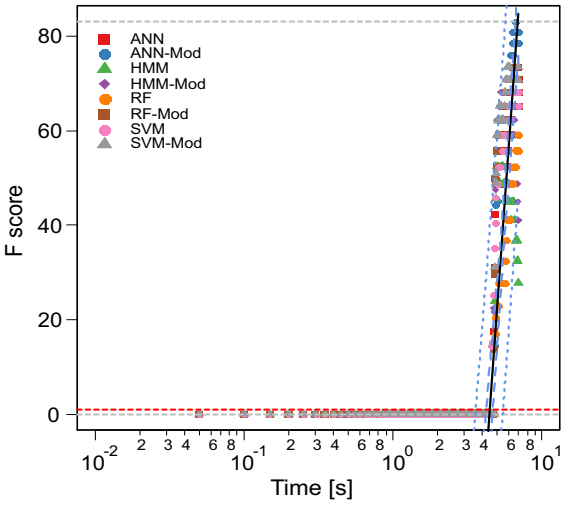

Figure 4.20: Right lane change classifier data [ADS20a]

The evaluation is premised on the classifier's prediction time as a function of remaining time before the event itself.

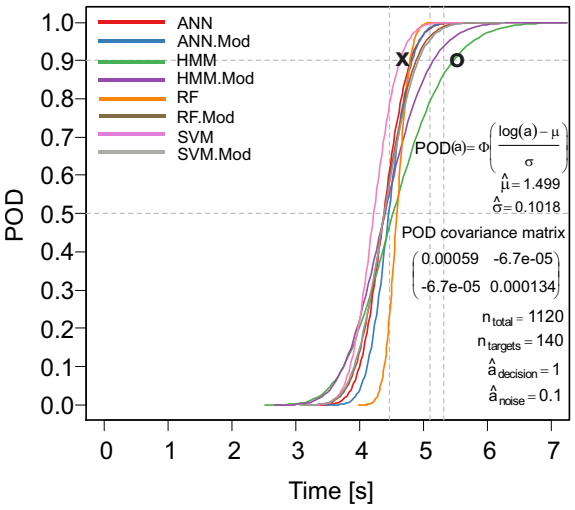

Figure 4.21: Left lane change POD results, **x**: best **o**: worst [ADS20a]

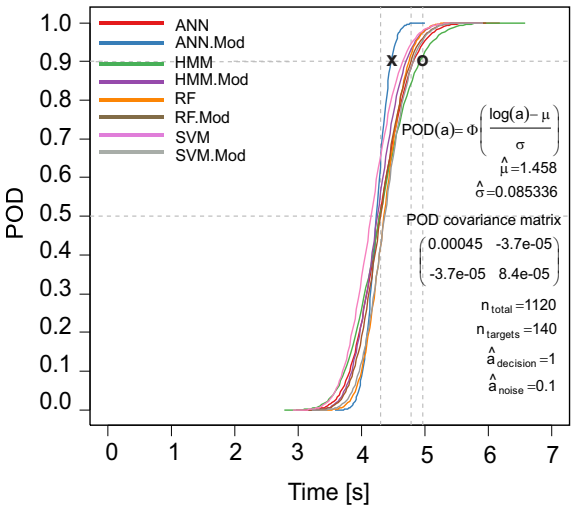

Figure 4.22: Right lane change POD results, **x**: best **o**: worst [ADS20a]

The POD curves of all classifiers are generated for left change (Fig. 4.21) prediction and right lane change (Fig. 4.22) prediction using F score as target response evaluation. There are currently many certification standards. The 90/95 reliability measure is used here. The 90/95 certification in this context expresses the time required to detect complete lane change with 90 % probability at 95 % confidence level. A classifier in this context is considered to be better if it has a lower 90/95 time value compared to another. The algorithm is able to predict the complete

lane change behavior faster so it has better prediction capabilities. The 90/95 POD values for left and right lane changes are illustrated in Tab. 4.2.

Table 4.2: POD results for all classifiers using FM response values

Classifier	Left lane POD [s]	Right lane POD [s]
ANN	4.959	5.011
HMM	5.776	5.138
RF	4.876	4.896
SVM	4.781	4.823
ANN-Mod	4.969	4.565
HMM-Mod	5.464	4.871
RF-Mod	5.076	4.952
SVM-Mod	5.113	4.979

For this specific example, the experimentally obtained and analyzed results can be summarized as follows:
In estimating left lane change classification results:

1 Conventional SVM produced best results (4.781 s) with conventional HMM producing worst results (5.776 s).

2 The use of prefilters and tuning parameters with the aim to optimize the performance of conventional algorithms did not result in improved POD, as three conventional classifiers (ANN, RF, and SVM) performed better than the modified models.

In estimating right lane change classification results:

1 Modified ANN produced best results (4.565 s) with conventional HMM producing worst results(5.138 s).

2 The use of prefilters to tune parameters with the aim to optimize the performance of conventional algorithms generally did result in improved POD, as three modified classifiers (ANN-Mod, RF-Mod, and SVM-Mod) performed better than the conventional models.

Accordingly we can conclude that the best classifier for this example task is conventional SVM for left lane change prediction while modified ANN is best for right lane change prediction. Conventional HMM produced worst results in both instances. An interesting observation from the results is that, the use of prefilters to tune parameters with the aim to optimize the performance of conventional algorithms is achieved for right lane change prediction whilst worst results are obtained for left lane change prediction.

4.4.1 Improvement of classifier detection capability

Comparison of the introduced method with improved approach is made in this section. From the concrete example applying the approach introduced it can be generalized that the generated POD is not unique but dependent on target response. To improve the POD results, a strategy involving the utilization of crisp target response evaluation value will be used. The FM is calculated using both the DR and precision (PPV). The DR depend solely on the test dataset whilst precision is dependent on the test and population. The DR response value can thus be selected and used for evaluation to produce a better POD though the FM provides a better generalization and not restricted to just the test data. The DR is used as response value and the 90/95 POD values for all eighth classifiers are shown in Tab. 4.3.

Table 4.3: POD results for all classifiers using DR response values

Classifier	Left lane POD [s]	Right lane POD [s]
ANN	3.067	3.325
HMM	1.439	1.204
RF	4.219	3.354
SVM	2.433	2.951
ANN-Mod	1.143	3.679
HMM-Mod	0.6355	0.5748
RF-Mod	2.883	3.181
SVM-Mod	3.300	3.485

The results show an improvement in the POD values for all classifiers in comparison to when FM was used as response value. This simple but effective strategy can be implemented to improve the POD results. It is also observed that usage of prefilters to tune parameters with the aim to optimize the performance of conventional algorithms generally did work with the modified classifiers producing better POD results. The introduced POD characterization strategy shows a strong dependence on the response value and hence consideration should be given when selecting the target response values for optimal classifier detection.

The procedure is extended to two other drivers to investigate the generalize performance of the individual classifiers. The results are shown in Tab. 4.4 and 4.5.

A summary of the classification results for all three drivers are:
In estimating left lane change classification results:

1 For driver 1: Modified HMM produced best results (0.6355 s) with RF producing worst results (4.219 s).

Table 4.4: Left lane change 90/95 POD [ADS20]

Algorithm	Driver 1 POD [s]	Driver 2 POD [s]	Driver 3 POD [s]
ANN	6.383	3.067	4.400
HMM	3.518	1.439	3.052
RF	4.680	4.219	5.840
SVM	2.548	2.443	1.182
ANN-Mod	5.611	1.143	3.128
HMM-Mod	2.801	0.6355	2.853
RF-Mod	5.185	2.883	5.223
SVM-Mod	4.084	3.300	5.412

Table 4.5: Right lane change 90/95 POD [ADS20]

Algorithm	Driver 1 POD [s]	Driver 2 POD [s]	Driver 3 POD [s]
ANN	3.325	5.605	4.568
HMM	1.204	3.352	3.593
RF	3.354	3.974	10.9
SVM	2.951	4.037	5.525
ANN-Mod	3.679	4.143	4.302
HMM-Mod	0.5748	2.431	1.625
RF-Mod	3.181	6.005	5.631
SVM-Mod	3.485	5.688	5.616

2 For driver 2: SVM produced best results (2.548 s) with ANN producing worst results (6.383 s).

3 For driver 3: SVM produced best results (1.182 s) with RF producing worst results (5.84 s).

In estimating right lane change classification results:

1 For driver 1: Modified HMM produced best results (0.5748 s) with modified ANN producing worst results (3.679 s).

2 For driver 2: Modified HMM produced best results (2.431 s) with modified RF producing worst results (6.005 s).

3 For driver 3: Modified HMM produced best results (1.625 s) with RF producing worst results (10.9 s).

Accordingly the following statements can be deduced for the experimental results:

1 For this example task, the most suitable classifier is modified HMM producing 4/6 best results.

2 The worst classifier for this example task is RF/ modified RF producing 4/6 worst results.

3 The application of prefilter to define features and influence prediction performance generally results in an improved POD except for SVM.

The introduced approach permits a new POD-based certification and comparison method for binary classifiers based on their reliability of prediction. Evaluation and prediction of an impending event seconds before it occurs by the applied eight classifiers is not directly possible using known measures (DR, FAR, ROC). However the implemented approach provides a suitable strategy in predicting future event of a dynamically changing situation as a function of the process parameter (here: time) seconds before it occurs. Using a typical 90/95 POD certification standard provides a new measure to compare the classification performance of different machine learning approaches to a given task.

4.4.2 Fusion of results

In this section, the robustness of the POD measure in detecting and predicting lane change behavior is improved through fusion. Robustness in this context refers to improved stability bounds. An option to access improved stability bounds is the tightness of confidence bounds. Three fusion approaches: Majority Voting, Bayesian Combination Rule, and Behavior Knowledge Space are implemented in this work to ascertain the possibility of improving the stability bounds. Also different static selection strategies are considered to verify combinations that lead to better POD results. Instead of considering and comparing the individual classification results, the fusion of decisions from the classifiers can lead to an improvement of reliability and/or robustness [TJGD08] [RKS19]. To avoid the combination of classifiers with low accuracy and high dependency in comparison to other classifiers, ensemble selection methods are used. The static selection method chooses all unseen patterns based on validation errors during training phase. This approach selects an ensemble of suitable classifiers for combination [ME15] [BJSO14]. In this contribution the results of the eight classifiers are selected using the accuracy. As a conclusion from the studies in [RS16], the selection of only the best classifiers does not always lead to the best results. In some cases, the combination of very good and medium or very good and worse classifiers can lead to a better fusion performance. Therefore the considered static selection strategies are

- all not modified ones,
- all modified ones,
- the two best,
- the three best,
- the four best, and
- the two worst.

The results of the classifiers are fused with the following three fusion methods using the classifier results in the form of class labels.

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

$$b_k(x \in c_i) = \begin{cases} 1, & \text{if } q_k(x) = c_i \\ 0, & \text{otherwise.} \end{cases} \tag{4.5}$$

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

$$b(x \in c_i) = \sum_{k=1}^{n_K} b_k(x \in c_i). \tag{4.6}$$

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

$$\hat{q}(x) = \begin{cases} c_i, & \text{if} \\ & b(x \in c_i) = \max\left\{b(x \in c_1), \ldots, b(x \in c_{n_C})\right\} \\ & \geq \alpha \cdot n_K + f_t(x) \\ 0, & \text{otherwise,} \end{cases} \tag{4.7}$$

where n_C denotes the total number of classes and $\alpha \in (0, 1]$ 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 4.7 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 $f_t(x) = 0$ [SL00] [RG00].

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

$$C^k = \begin{bmatrix} C^k_{11} & C^k_{12} & \dots & C^k_{1n_C} \\ C^k_{21} & C^k_{22} & \dots & C^k_{2n_C} \\ \vdots & \vdots & \ddots & \vdots \\ C^k_{n_C1} & C^k_{n_C2} & \dots & C^k_{n_Cn_C}, \end{bmatrix}, \tag{4.8}$$

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

$$P^k_{ij} = P(x \in c_i \mid q_k(x) = c_j) = \frac{C^k_{ij}}{\sum_{i=1}^{n_C} C^k_{ij}}. \tag{4.9}$$

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

$$P^k = \begin{bmatrix} P^k_{11} & P^k_{12} & \dots & P^k_{1n_C} \\ P^k_{21} & P^k_{22} & \dots & P^k_{2n_C} \\ \vdots & \vdots & \ddots & \vdots \\ P^k_{n_C1} & P^k_{n_C2} & \dots & P^k_{n_Cn_C} \end{bmatrix}. \tag{4.10}$$

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

$$bel(x \in c_i) = \frac{\prod_{k=1}^{n_K} P^{n_K}_{iq_k}}{\sum_{i=1}^{n_C} \prod_{k=1}^{n_K} P^{n_K}_{iq_k}}, \tag{4.11}$$

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 $BKS(q_1(x), \ldots, q_{n_K}(x)) = BKS_l$, where $l = 1, \ldots, n_C^{n_K}$. 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

$$c_l = \{c_i \mid n_l(c_i) = \max\big\{n_l(c_1), \ldots, n_l(c_2)\big\}\}. \tag{4.12}$$

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

$$\hat{q}(x) = \begin{cases} c_l, & \text{if } n_l > 0 \text{ and } \dfrac{n_l(c_l)}{n_l} \geq \lambda \\ 0, & \text{otherwise.} \end{cases} \tag{4.13}$$

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].

4.4.3 POD generation of fused data

Based on the computed values, the target response method is utilized in this section to generate POD for the classification/fused results as a new reliability standard. The aim is to establish a POD characterization to illustrate the effect of process parameters (here: time) to the classification/fused results so the probability distribution is derived.

Four models comprising combinations of logarithmic and linear scales are established for each fused ML algorithm to ascertain best model [MHA09].

The model that fits the mentioned criteria best is selected. Regression analysis is implemented on the selected model and the decision threshold is constructed as illustrated in Fig. 4.23.

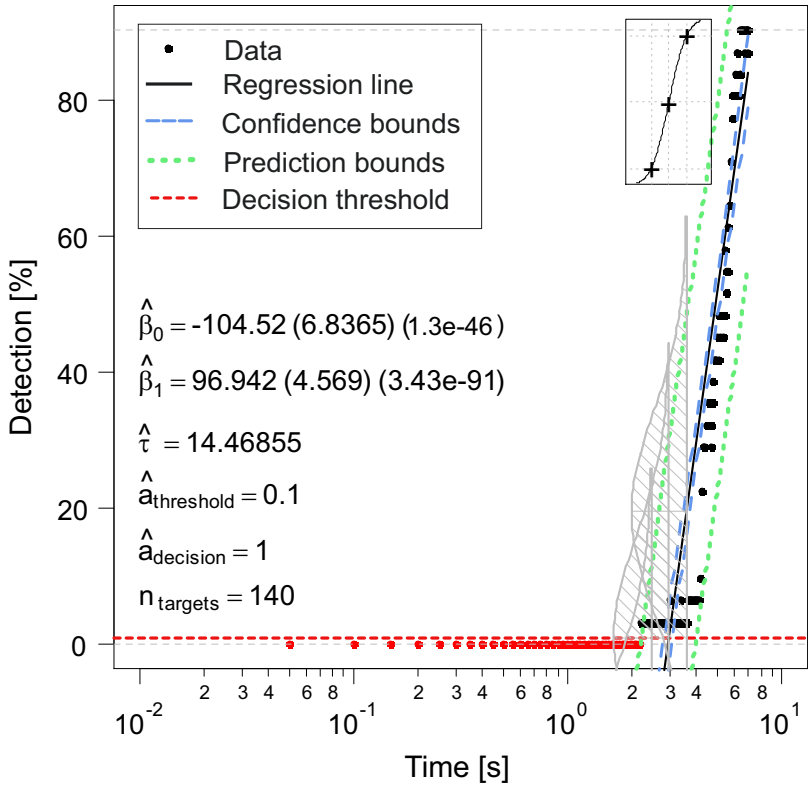

Figure 4.23: POD generation procedure

The POD curve is generated using area of the cumulative density function above decision threshold. The POD for the classifiers are constructed for both right and left lane change.

As example, the POD curve for ANN right lane change and 2 best BKS fused results are shown in Fig. 4.24 and 4.25 respectively.

4.5 Results and discussion

Detailed analysis of the results is discussed in this section. Based on the proposed approach, the POD curves of all classifiers are generated. There are currently many certification standards but the 90/95 reliability measure is used here. The 90/95 certification in this context expresses the time required to detect complete lane change with 90 % probability at 95 % confidence level. A comparison of Fig. 4.24 and 4.25 indicate the POD curve for the fused results has improved stability bounds due to its tight confidence bounds. The tight confidence bounds provides the desired robust lane change detection capabilities.

Classifiers are selected based on the mean overall accuracy with the ranking: The

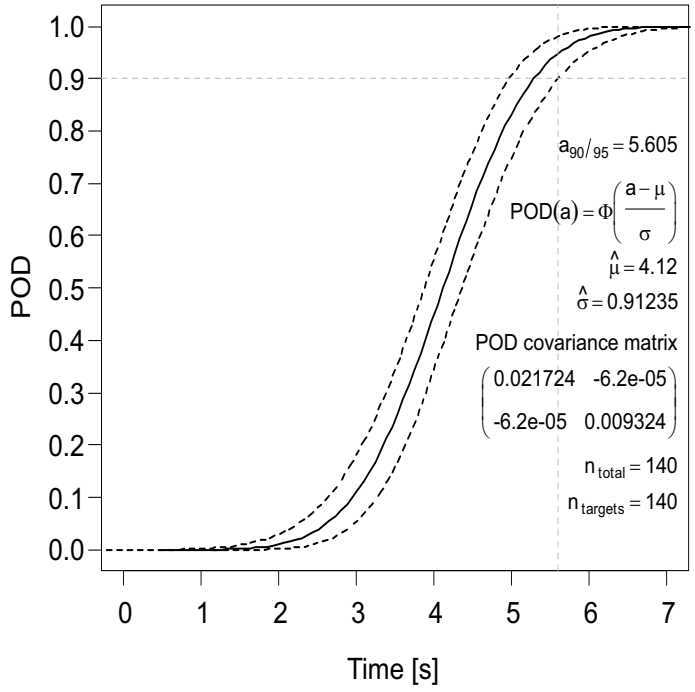

Figure 4.24: ANN right lane change

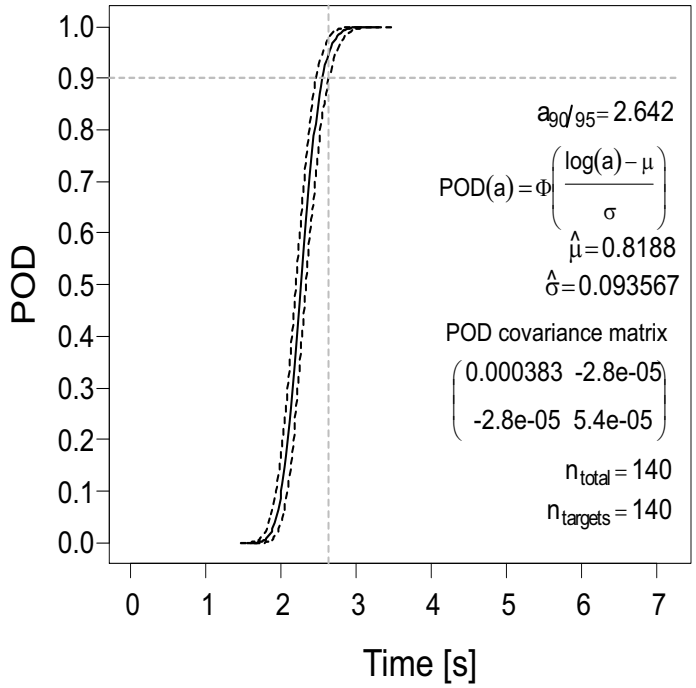

Figure 4.25: POD for 2 worst MV results

POD results for the different selection strategies are indicated in Tab. 4.6 and 4.7.

1. SVM-Mod 86.22 %
2. RF-Mod 85.90 %
3. ANN-Mod 84.85 %
4. RF 83.40 %
5. ANN 81.39 %
6. HMM-Mod 81.25 %
7. HMM 74.91 %
8. SVM 73.90 %

Table 4.6: Left lane change 90/95 POD

	Fusion approach		
Fused classifiers	BCR	BKS	MV
All not modified	-	3.276	3.271
All modified	-	3.333	2.649
2 best	-	3.315	3.107
3 best	3.765	3.141	3.234
4 best	-	-	3.544
2 worst	-	2.642	2.642

Table 4.7: Right lane change 90/95 POD

	Fusion approach		
Fused classifiers	BCR	BKS	MV
All not modified	4.192	3.398	3.303
All modified	4.169	3.86	3.452
2 best	3.618	3.618	3.284
3 best	4.169	3.592	3.457
4 best	4.509	4.045	3.475
2 worst	3.382	3.382	1.321

The time duration for complete lane change is 7 s. The least time values represent best results because the selected fused algorithm is able to predict the lane change within the shortest possible time.

The experimentally obtained and analyzed results can be summarized as follows: In estimating left lane change classification results for the fused data:

1 Majority voting is best suited producing 5/6 best results.

2 Bayesian combination rule has worst suited 6/6 worst results.

3 Two worst selection strategy produced best POD results (2.642 s)

In estimating right lane change classification results for the fused data:

1 Majority voting is best suited producing 6/6 best results.

2 Bayesian combination rule has worst suited 4/6 worst results.

3 Two worst selection strategy produced best POD results (1.321 s)

Accordingly the following statements can be deduced for the experimental results:

1 For this example task, the most suitable fusion approach is MV.

2 The worst fusion approach is BCR.

3 The best static selection strategy is 2 worst classifiers.

To understand the reason for the two worst classifiers producing best results requires knowledge on the POD results for the individual classifiers illustrated in Tab. 4.3. The results from Tab. 4.3 indicates that HMM-Mod has the best driver behavior prediction though SVM-Mod has the highest accuracy. An indication that an ML approach with the highest accuracy does not translate into it producing the best POD.

4.6 Summary

In this chapter a new assessment on the performance of classifiers using POD approach is presented. This is needed because the evaluation process of known diagnostic tools like ROC/PR curves are non-parametric and hence incapable to relate detectability with a process parameter. It can be clearly shown that for a selected decision threshold there is a corresponding unique FAR but several DR values. This implies, the ROC/PR curve though successful then suffers serious deficiencies in modern systems and applications which requires detectability to be related to a parameter. This deficiency is addressed using the new established POD measure. The new approach is demonstrated on experimental data from real human driving simulator. The target response method is implemented as a new analysis and certification tool for classifiers permitting the comparison of different ML algorithms. The introduced approach provides a comprehensive interpretation of the quality of a classification model by incorporating a process parameter in the evaluation process. Different fusion strategies are implemented to improve the reliability of POD statements through enhancement of POD stability bounds. The target response analysis is used to compare different fusion and static selection strategies. The fusion approaches used are Bayesian Combination Rule, Behavior Knowledge Space, and Majority Voting. The results show MV is the best fusion approach for this example task in estimating lane change behavior. The introduced approach allows a new insight into the ability of machine learning-based classifiers to handle dynamically changing situations.

5 Implementation of POD as new evaluation measure for vision-based classification approaches

In this chapter, object classification and detection using Deep Convolutional Network (CNN) is presented. Deep CNN is a useful classifier for large-scale image classification. It is common knowledge that classifier performance depends on problem characteristics and vary with process parameters. The question about the perfect classifier is difficult to evaluate. This chapter focuses on accurate and explicit evaluation of CNN classifiers in image detection. Existing evaluation approaches do not integrate the effects of other properties related to image parameters or varying problem details in the evaluation process. The POD is proposed as a new evaluation measure allowing evaluation of further image parameters affecting the classification results. The introduced approach permits the assessment and comparison of different classifiers related to the same parameter-dependent classification task. The proposed method is implemented on a number of images and comparisons made on parameters with the best detection capabilities. Similar to typical standards in the industry, the probability of detecting an image at 90 % with a reliability level of 95 % is successfully implemented and the comparison made between different image parameters. The advantage of the novel approach is experimentally evaluated for vision-based classification results of CNN approach considering different image parameters. To illustrate the newly introduced performance measure a comparison using alternative standard machine learning-based classifiers using identical experimental data is shown. This chapter is organized as follows: In section 5.1 the state-of-art and the need for a new evaluation approach is presented. In section 5.2 the classification approaches and the corresponding results are discussed. In section 5.3 newly proposed approach [ADS19] based on POD measure is briefly introduced, applied, and the results are presented. Comparison is made between different classifiers in section 5.4 using Hit/Miss approach and subsequently the conclusion is presented in section 5.5.

The content, figures, and tables in this chapter are based on the submitted paper [BAMS20].

5.1 Introduction

Performance evaluation and analysis of classification approaches are based on either a numeric measure or a graphical representation. The numeric measures are based on the calculation of true positives, false positives, false negatives, and true negatives. They can be illustrated in a confusion matrix as a basic performance metric [Sus04]. Selection of a suitable numeric measure normally depends on the classification purpose. In [SL09] twenty four performance measures are introduced

for binary, multi-class, multi-labeled, and hierarchical classifiers based on the values of the confusion matrix. The confusion matrix evaluations used are accuracy, precision, sensitivity, recall, fscore, exact match ratio, among others. The limitation of all aforementioned measures are that some important details could be hidden by using a single performance measure. These performance measures are also sensitive to imbalanced data. Nevertheless using one scalar performance measure is very simple and easy. Accuracy measure (as a ratio between the correctly classified samples to the total number of samples) is the most common performance measure in evaluating classification algorithms. In [LJL+19] for example, a group collaborative representation model for face recognition and object categorization tasks are evaluated using the accuracy. In [WFY15] average precision (AP) and mean AP (mAP) are used in the evaluation of collaborative linear coding approaches with the purpose of eliminating the negative influence regarding noisy features in classifying images.

Graphical assessment methods provide different interpretations of the classification performance. The ROC is one of the graphical performance measures for binary classifiers [Faw06]. It is the illustration of true positive rate against false positive rate. This illustration is not helpful in the case of data sets with large class imbalances. An alternative visualization is the precision-recall (PR) curve that is the relationship between recall and precision. The area under the ROC curve (AUC) and PR curve (AUPRC) are also used as performance measures with the same limitations as single performance measures. In [DG06] a comparison between ROC and PR performance illustrations is investigated. Furthermore, some approaches are proposed to find a suitable threshold value to reduce the sensitivity of assessment to the imbalanced data [ZXL+16]. Another graphical assessment approach which shows the relation between false acceptance rate (FAR) and false rejection/recognition rate (FRR) is the Detection Error Trade-off (DET) curve [MDK+97].

Deep learning approaches are widely used in computer vision for recognition and tracking [LOW+20, ZZXW19]. For large-scale image classification, Deep CNN is one of the suitable classifiers . Previously mentioned classification assessment approaches can be used to evaluate the classification results of CNN in machine vision field [HZRS15]. The most commonly used performance measure in object detection field is AP considering both precision and recall into account [LOW+20].

It is worth mentioning that the effects of other properties related to image parameters or varying problem details are not considered in the well-known evaluation approaches. This limitation in verifying and incorporating the effect of image parameters on the classifier performance is addressed in this chapter. A new performance

evaluation is proposed based on the Probability of Detection (POD) evaluation metric for the evaluation and comparison of different binary classifiers.

In [ADS19] the authors introduced the idea of classifier evaluation using POD and applied this to driving behavior prediction. In this chapter the focus is different: the approach is applied to vision-based image detection, the theory of the new approach is completely developed, and a new certification standard for image classification approaches is clearly defined.

5.2 Deep Convolutional Neural Network for object classification

Convolutional Neural Network (CNN) is a class of deep Neural Networks and widely used in computer vision as a suitable approach for image recognition and classification. The most common advantage of using CNN compared to other image classification algorithms is the pre-processing filter that can be learned by the network. The structure of CNN consists of convolutional layers, ReLU layers (activation function), pooling layers, and fully connected layer. Each layer involves linear and/or nonlinear operators.

Convolutional layers extract features from the input image by applying a convolution operation to the input image. The convolutional layer is the first filter layer that builds a convolved feature map for the next layer. The pooling layers are applied to avoid overfitting and preserving the desired features. The last Pooling Layer is followed by one (or more) fully-connected layer(s) according to the architecture of the multilayer perceptron. This is mainly used for classification. The number of neurons in the last layer then usually corresponds to the number of (object) classes the network should distinguish.

With each filter layer the abstraction level of the network increases. The abstraction that finally leads to the activation of the subsequent layers is determined by the characteristic features of the given classes that will be recognized.

5.2.1 Network architecture

In this chapter, two architectures are used as the classification model:

1. ResNet18 [HZRS16] and

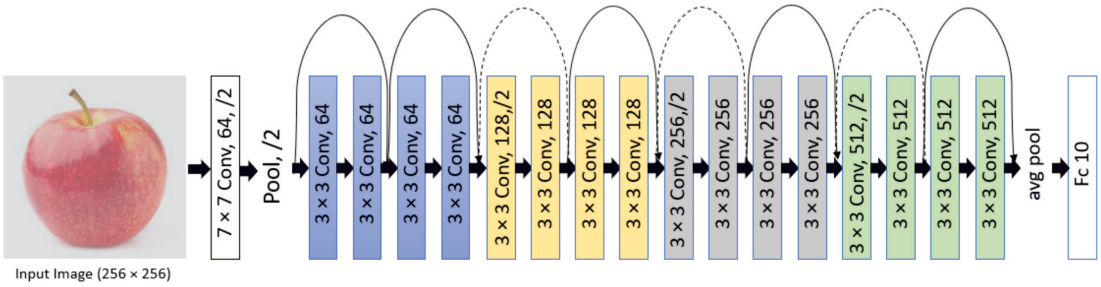

Figure 5.1: Architecture of ResNet18 (according to [HZRS16])

2. MobileNet V2 [SHZ+18].

In the ResNet18 structure, a CNN with 18 layers deep is built based on a deep residual learning framework shown in Fig. 5.1. The learning approach eliminates accuracy degradation as the network depth increases. In the MobileNet V2 structure, an inverted residual and linear bottleneck layer is proposed to increase the accuracy in mobile and computer vision applications. This structure is an improved version of MobileNet V1 which uses the depthwise separable convolution to decrease the size of the model (number of parameters) and correspondingly the complexity (multiplications and additions) of the network. Overall architecture of this approach is shown in Fig. 5.2 and correspondingly in Table 5.1.

Table 5.1: Overall architecture of MobileNet V2

Input	**Operator**	t	c	n	s
224^2 x 3	conv2d	-	32	1	2
112^2 x 32	bottleneck	1	16	1	1
112^2 x 16	bottleneck	6	24	2	2
56^2 x 24	bottleneck	6	32	3	2
28^2 x 32	bottleneck	6	64	4	2
14^2 x 64	bottleneck	6	96	3	1
14^2 x 96	bottleneck	6	160	3	2
7^2 x 160	bottleneck	6	320	1	1
7^2 x 320	conv2d 1x1	-	1280	1	1
7^2 x 1280	avgpool 7x7	-	-	1	-
1 x 1 x 1280	conv2d 1x1	-	k	-	-

5.2.2 Classification results

Custom dataset containing 10 classes (apple, atm, card, cat, banana, bangle, battery, bottle, broom, bulb, calender, camera) is used in this work. Both ResNet and

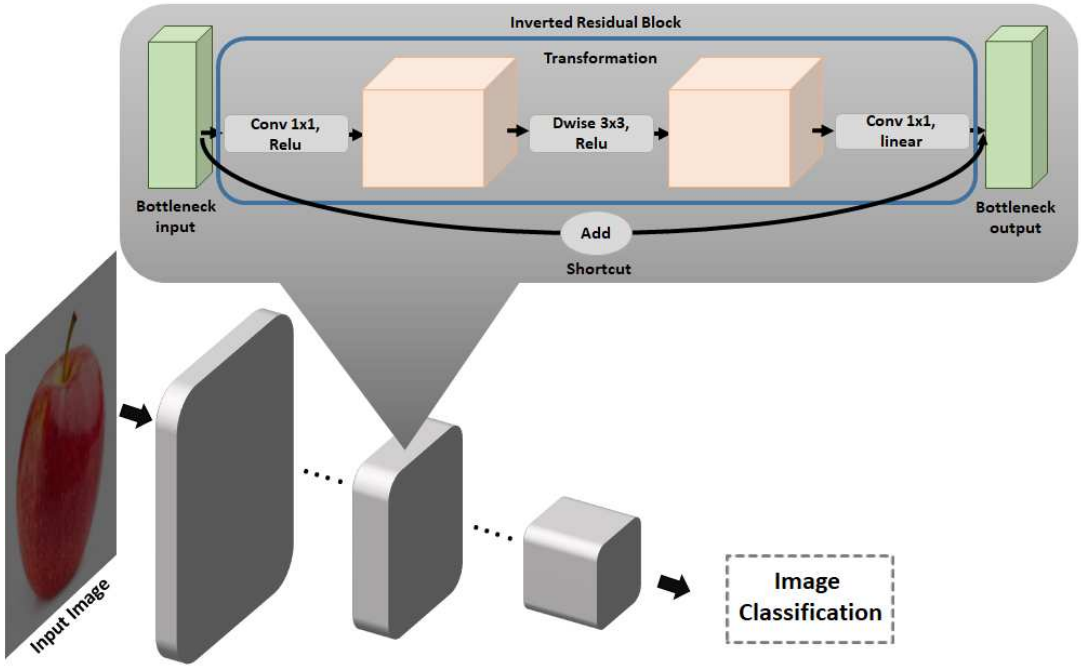

Figure 5.2: Network architecture of MobileNet V2 with basic building block as inverted residual block (according to [SHZ+18])

Table 5.2: Classification results

Model	Training accuracy	Validation accuracy
ResNet18	0.995	0.9878
MobileNet V2	0.9621	0.976

MobileNet V2 are trained for 10 epochs on the customized data set using transfer learning approaches [TSK+18]. The pre-trained version of ResNet and MobileNet V2 on the ImageNet dataset [DDS+09] is used in the experiments. It can be observed that it is faster to achieve high classification accuracy on the customized dataset using pre-trained models trained on almost similar datasets. The classification results are detailed in Table 5.2.

5.3 New evaluation of object classification by the target response approach

In this research, the two classifiers are used in detecting an apple image. The varying image parameters considered are the contrast and brightness (see Fig. 5.3). The target response approach is adapted and implemented as a new performance

Figure 5.3: Input images with varying brightness (top row) and varying contrast (bottom row) provided to the classification networks

Table 5.3: Probability of detection considering different image contrasts

Contrast	Probability
0.0	0.31452537
0.1	0.72617805
0.2	0.9944179
0.3	0.9994198
0.4	0.99984
0.5	0.9999324
0.6	0.99995387
0.7	0.99997306
0.8	0.99997616
0.9	0.9999759
1.0	0.9999771

assessment for object detection classification models. For illustrative purpose, the MobileNet V2 classifier and contrast values are implemented. The contrast is varied from 0 (min) to 1 (max). Beyond a contrast value of 0.2, the algorithm is able to detect the image with a probability of 0.9944 (see Table 5.3). Therefore the analysis is made for the range of values 0-0.2 since beyond 0.2, the algorithm reaches saturation threshold and has no effect on POD characterization. One hundred contrast values between 0-0.2 and the corresponding probability values are examined.

A graph of the MobileNet V2 classifier contrast data is illustrated in Fig. 5.4. The least detectable target (inspection threshold), maximum detectable target (saturation threshold), and threshold below which the observed data is characterized as

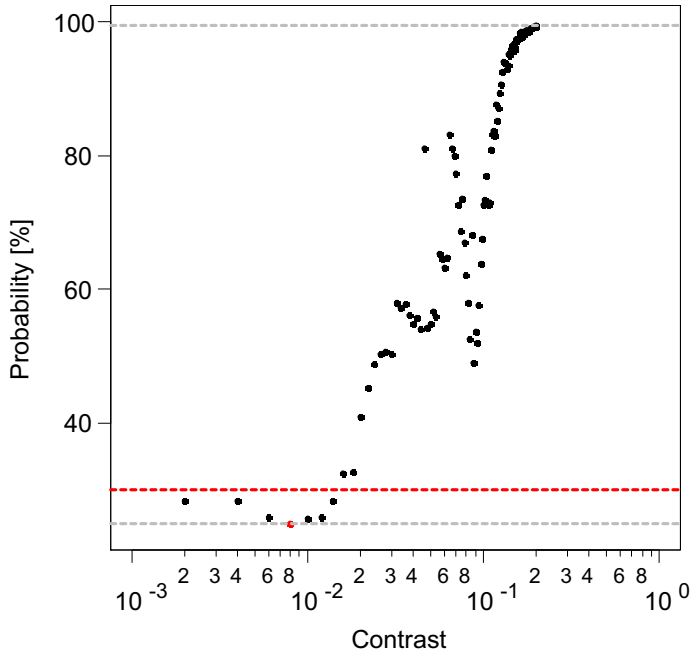

Figure 5.4: Log x data model

noise(decision threshold) are constructed.

Regression analysis are carried out for the dataset to establish a POD characterization to illustrate the effect of process parameter (here: contrast) to the detection probability results. The 95 % confidence bounds and the prediction bounds are also constructed (see Fig. 5.5). The confidence bounds are constructed for certification purpose while the prediction bounds serve as boundaries so that for every new 100 observations, 95 should fall within. The cumulative density function (CDF) for the data distribution are also constructed. The POD curve is generated using area of the cumulative density function above decision threshold (Fig. 5.6).

The confidence bounds about the regression line are used in constructing the 95 % bounds around the POD curve. The POD generation procedure are repeated for the confidence bounds to construct the 95 % bounds around the POD curve. The 90/95 certification criteria representing 90 % probability of detecting the image with a reliability of 95 % is utilized in this research. The 90/95 (**o** in Fig. 5.7) value for this concrete example is 0.02167. This implies, in varying the contrast values from 0 to 1, at 0.02167 the algorithm is able to correctly detect the image with 90 % probability at 95 % reliability. Unlike classical evaluation approaches, this introduced and novel approach provides a new evaluation and performance assessment for classifiers incorporating the effect of image property on the classification results.

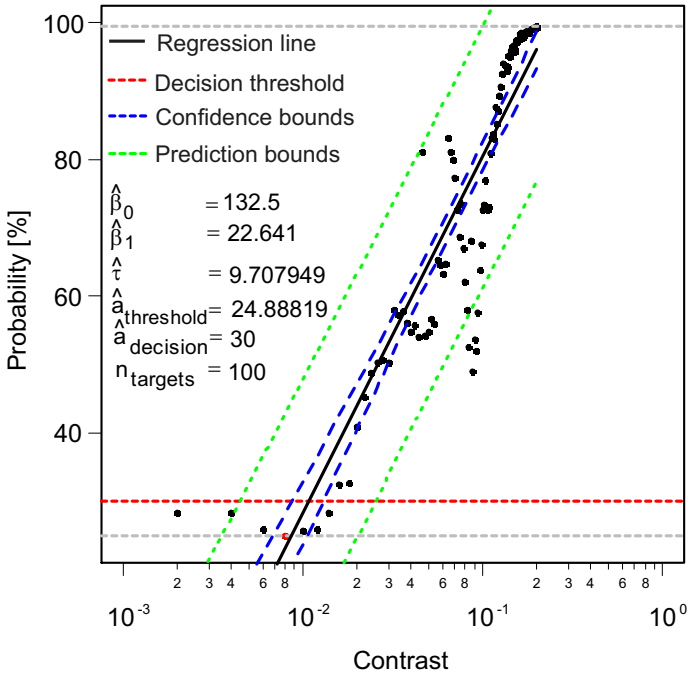

Figure 5.5: Regression analysis

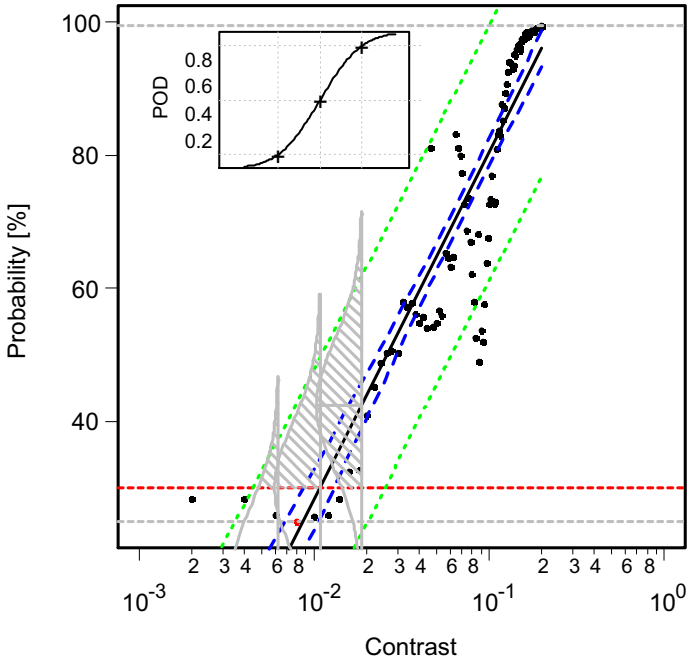

Figure 5.6: POD generation approach

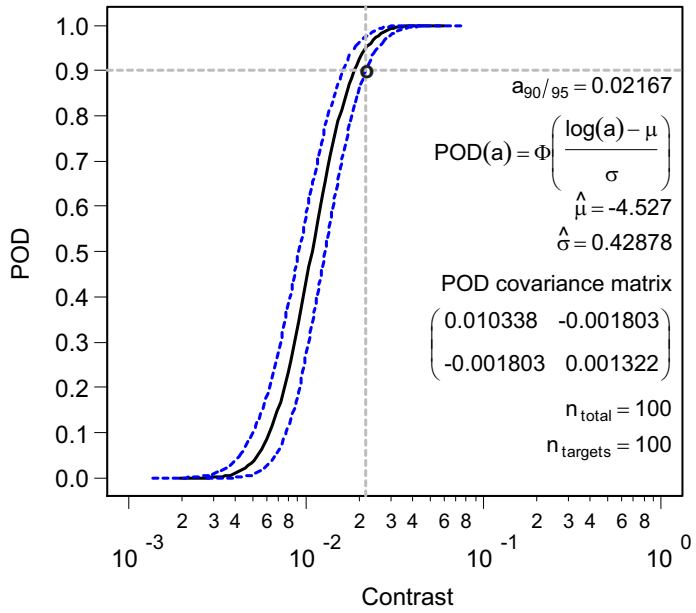

Figure 5.7: POD curve, **o**: 90/95 POD value

5.4 Comparison of different Neural Network models by the Hit/Miss approach

In this section a new POD implementation strategy is deployed due to the nature of the data to be analyzed. This new approach is used to evaluate the performance of two Neural Network algorithms namely MobileNet V2 and ResNet. Each model is implemented on two varying parameters namely brightness and contrast. Here, the classes as opposed to the probabilities are used. The classes are binary in nature and hence the Hit/Miss approach will be used. At each parameter intensity, if the algorithm correctly classifies the image it is assigned a value of 1 (Hit) and 0 (Miss) for a wrong class. Like the former approach the analysis is made for parameter changes from 0-0.2 at a step size of 0.002. To accurately implement the Hit/Miss approach, the appropriate link function to be used needs to be determined. For illustrative purpose, the classification results of MobileNet V2 considering brightness values are used (Fig. 5.8).
Eight models of different link functions are constructed to select best fitting model as illustrated in Fig. 5.8. The data distribution of the logarithmic model does not support the construction of a valid POD curve estimate and hence the Cartesian models will be more suitable for this example. A statistical hypothesis testing procedure is undertaken to ascertain best fitting link function model. To model the relationship of GLM, deviance is a measure of goodness of fit: the smaller the deviance, the better the fit. It is attained using a generalization of the sum of squares of residuals in ordinary least squares to cases where model-fitting is achieved by maximum likelihood. The Cartesian Weibull is selected for this specific analysis because it has

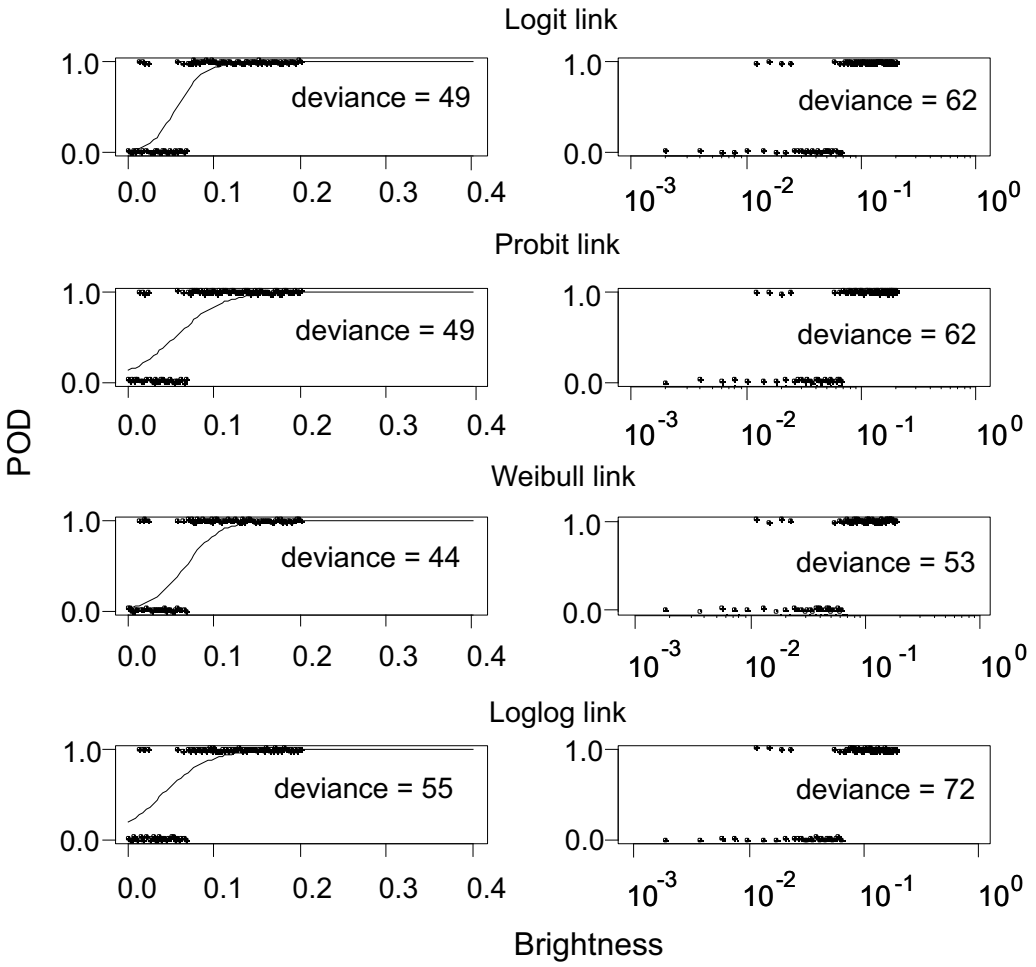

Figure 5.8: Models to select link function with least deviance [BAMS20]

the least data deviance in comparison to the other link functions. The procedure is repeated for MobileNet V2 contrast parameter as well ResNet brightness and contrast parameters. The hypothetical testing procedure show Cartesian Weibull is the best link function for all. The Weibull is implemented to map $-\infty < x < \infty$ to $0 < y < 1$. The Weibull function is expressed as

$$f(X) = g(y) = log(-log(1-p)),$$

where $f(X)$ is an algebraic function with linearized parameters and p the probability. The probability of detection as a function of brightness B for the Weibull model is

$$POD(B) = 1 - exp(-exp(f(X))).$$

Using the Weibull, surface contours are constructed to ascertain the most plausible GLM (Fig. 5.9, 5.10, 5.11, and 5.12).

The likelihood ratio test is used to assess goodness of fit. The log likelihood ratio contour encloses all β_0, β_1 pairs that are plausibly supported by the data. The

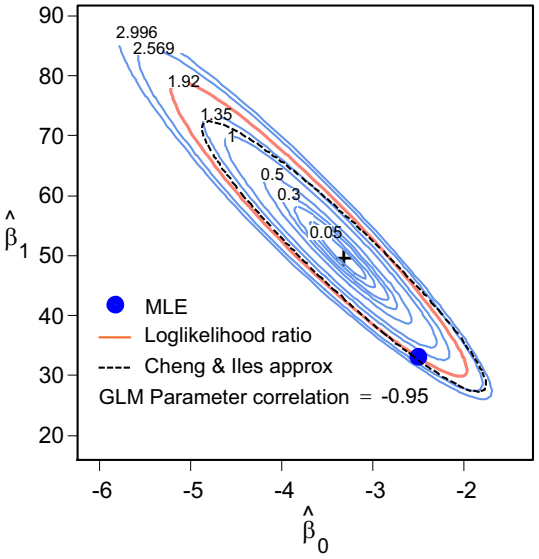

Figure 5.9: Likelihood surface contour for MobileNet V2 brightness [BAMS20]

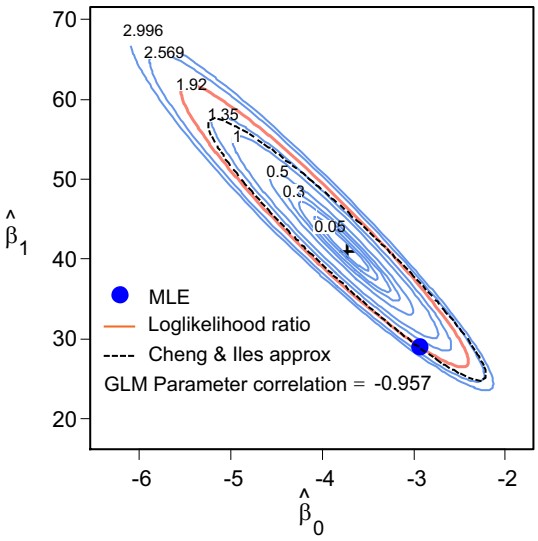

Figure 5.10: Likelihood surface contour for MobileNet V2 contrast [BAMS20]

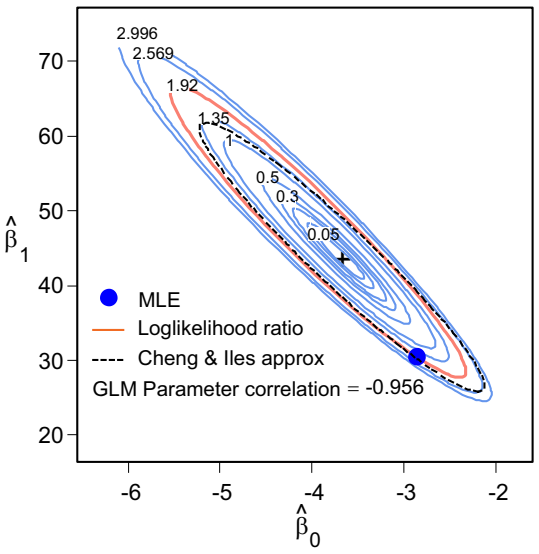

Figure 5.11: Likelihood surface contour for ResNet brightness [BAMS20]

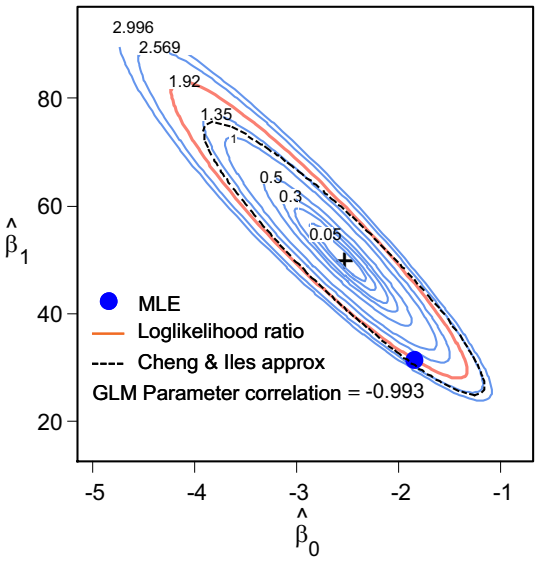

Figure 5.12: Likelihood surface contour for ResNet contrast [BAMS20]

confidence bounds are constructed on the surface contours using a method developed by Cheng & Iles approximation [CI83].

Using the maximum likelihood estimation (MLE) method, the GLM values for the intercept and gradient are generated for both classifiers. With these values, a GLM model and confidence bounds fitting the data are constructed (Fig. 5.13, 5.14, 5.15, and 5.16).

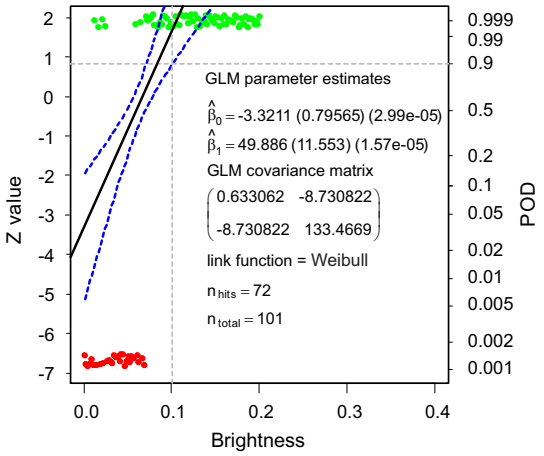

Figure 5.13: GLM with estimated parameters for MobileNet V2 brightness [BAMS20]

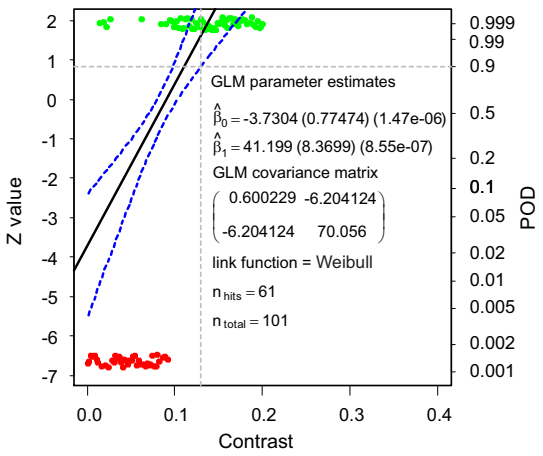

Figure 5.14: GLM with estimated parameters for MobileNet V2 contrast [BAMS20]

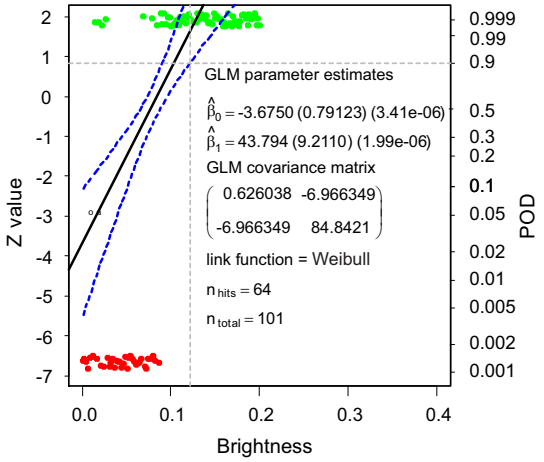

Figure 5.15: GLM with estimated parameters for ResNet brightness [BAMS20]

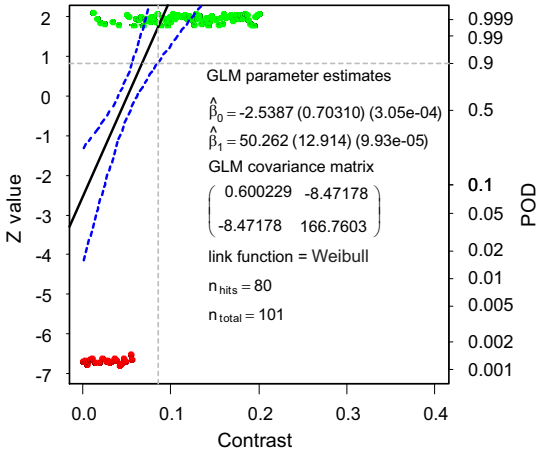

Figure 5.16: GLM with estimated parameters for ResNet contrast [BAMS20]

From the fitted GLM model, the POD curve is generated using the parameter estimates (Fig. 5.17, 5.18, 5.19, and 5.20). The 90/95 criteria is also successfully implemented on the results.

The 90/95 reliability value for MobileNet V2 is 0.101. This implies at the exact brightness intensity value of 0.101, the MobileNet V2 algorithm is able to detect the image with 90 % POD at a 95 % reliability level. The procedure is repeated for MobileNet V2 contrast, ResNet brightness, and ResNet contrast parameters. A comparison of the 90/95 POD values for the two classifiers and image parameters are detailed in Table 5.4.

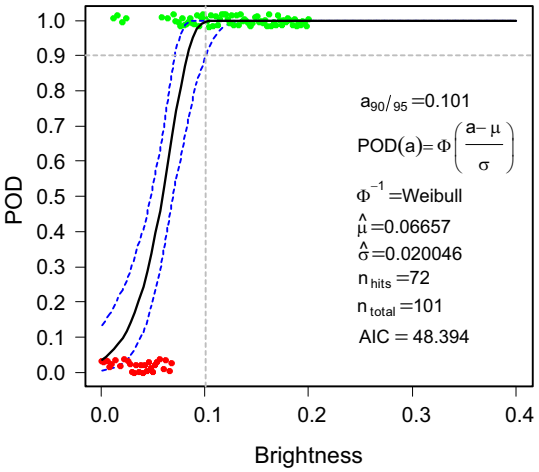

Figure 5.17: POD curve for MobileNet V2 brightness [BAMS20]

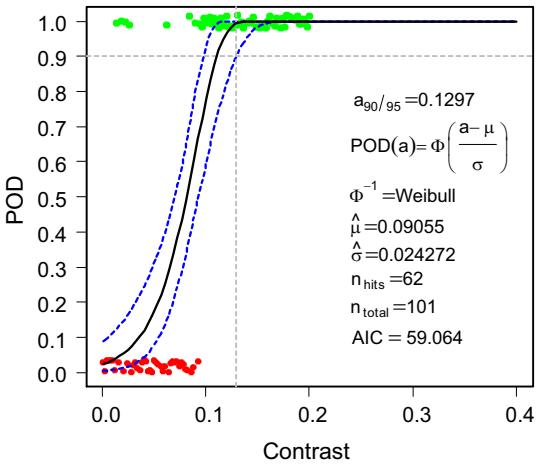

Figure 5.18: POD curve for MobileNet V2 contrast [BAMS20]

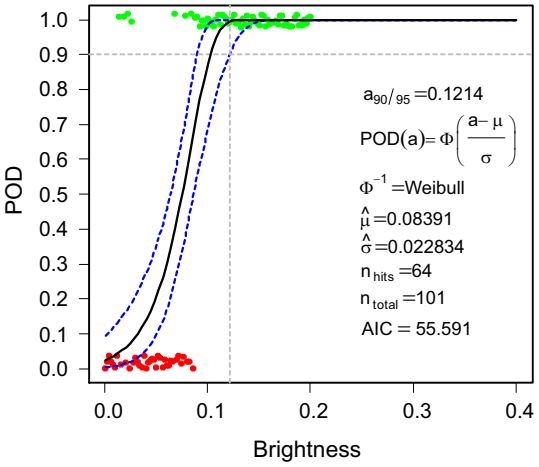

Figure 5.19: POD curve for ResNet brightness [BAMS20]

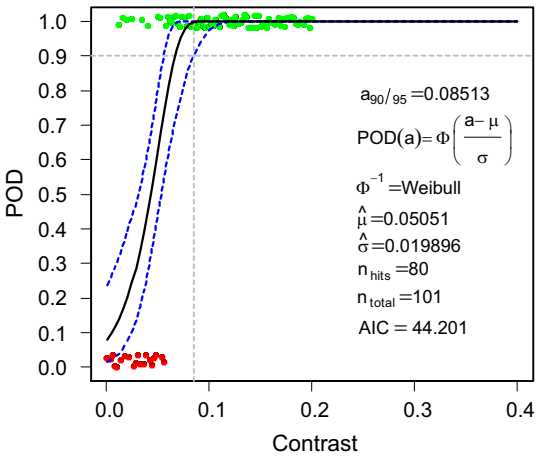

Figure 5.20: POD curve for ResNet contrast [BAMS20]

Table 5.4: The 90/95 POD values [BAMS20]

Classifier	Brightness	Contrast
MobileNet V2	0.1010	0.1297
ResNet	0.1214	0.0851

The smallest 90/95 POD values define the best results because the classification algorithm is able to detect the image with 90/95 reliability with the least contrast/brightness changes. The results indicate Mobile2Net V2 performed better using brightness as parameter while ResNet performed better using contrast as parameter.

This novel POD approach introduces a new performance evaluation for vision-based classification models incorporating and assessing directly the effects of image parameters.

The introduced approach makes it possible to check the number of target hits and missed directly from the POD curve. For the classification results, there were no misses beyond 90/95. The hits beyond 90/95 for the classifiers and different parameters are detailed in 5.5.

Table 5.5: Hits beyond 90/95 POD [BAMS20]

Classifier	Brightness	Contrast
MobileNet V2	50	36
ResNet	40	58

For this specific example, the evaluation and prediction system is reliable for all classification results because no misses are observed beyond the 90/95 POD threshold. For this specific example, the classifiers with the most hits beyond 90/95 are the best because the most correct class predictions beyond 90/95 thresholds are generated. This assertion is corroborated by the POD results in Table 4. However this conclusive statements can only be made for systems with no misses beyond the 90/95 threshold. The complexity increases if there are misses beyond 90/95 and requires the performance of noise analysis to conclude on the best classifier.

5.5 Summary

This chapter presents a new measure for performance evaluation of binary classifiers used in object detection. This is needed because often the effects of other properties related to image parameters or varying problem details are not considered in the evaluation process. Effectiveness of this approach is experimentally evaluated for vision-based classification results of CNN considering different image parameters. Two classification models are trained (ResNet and MobileNet V2). The proposed target response approach is implemented on MobileNet V2 classifier as a new evaluation procedure. Furthermore, the results of two classifiers are compared using the novel Hit/Miss evaluation approach. The results indicate that Mobile2Net V2 performs better considering brightness as parameter while ResNet performs better considering contrast as varying parameter. Based on the newly introduced POD-related evaluation, different vision-based classification approaches can be clearly

distinguished with respect to their ability to predict the correct image as a function of related image parameters. The introduced approach provides a comprehensive interpretation of the quality of a classification model and therefore allows a new evaluation quality.

6 Summary, conclusion, and future work

6.1 Summary and conclusion

In this thesis challenges associated with monitoring dynamical systems are presented. The difficulties are related to data acquisition, feature extraction, and statistical modeling for features classification. A data acquisition system consists of software (signal data) and hardware (sensors and other components). The number, types, and sensor positions has been a difficult subject for researchers over the years. Feature selection is also critical due to different sensitivity levels of features to faults. Fault diagnosis approaches try to solve these problems however critical questions still remain: transition from theory to practical implementation, correctly selecting senors (types, number, and positions) and reliability evaluation of FDI schemes.

To address these challenges, a novel POD-oriented view to vibration-based diagnosis typically considering different sensor types is presented. The measurements used are acceleration, strain, and displacement (laser sensors). The results indicate that the POD characterization depends on the sensor type, sensor position, fault position, and the feature selected. The sensor type has an effect on the POD due to the fact that performance specifications vary for different sensors. Sensor and fault position plays an important role due to observable effects. The $a_{90/95}$ criteria representing probability of 90 % at a confidence level of 95 % is successfully implemented in vibration-based FDI as a reliability measure. The POD measure is adapted and implemented in this work to provide a statistical modeling technique to select the best sensor (type and position)-feature combination for optimal fault diagnosis especially considering their noise characteristics.

A new data driven noise analysis procedure is presented to analyze the PFP and corresponding POD value for any selected decision threshold. An effective method to improve sensor detection without affecting the false alarm by utilizing sensitive features is introduced. The noise analysis procedure provides a graphical representation that illustrates the diagnostic capabilities of a sensor as its decision threshold is varied. The new insight introduced allows task-/application-specific quantification of sensors relative to vibration-based monitoring/diagnosis of faults.

Using the novel 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. Noise analysis allows to define a decision threshold which permit a suitable trade-off between the POD and PFP.

The introduced approach is extended to the certification of ML approaches. This is needed because existing evaluation processes and known measures like ROC often do not consider effects of process parameters on classification results. When

interpreting scenarios or for classification of complex sensor data describing a complex spatial scenario this might be of special interest. The 90/95 POD reliability measure is implemented as a new certification standard for classifiers. The new approach is demonstrated on experimental data from real human driving simulator. The data from the multi-sensor system are fused by classification and subsequently decision/statements about the lane change behavior have to be made. The approach also permits the comparison of different ML algorithms thereby aiding in the selection of desirable approach for a specific task. The target response analysis is used to compare different classifiers. A new measure establishing a process-parameter dependent evaluation is introduced allowing a new insight into the ability of machine learning-based classifiers to handle dynamically changing situations.

To demonstrate the generalizable capabilities of the introduced approach and also to solve a critical problem in vision-based evaluation approaches, the proposed approach is extended into the field of computer vision systems albeit different implementation strategies. This is needed because often the effects of other properties related to image parameters or varying problem details are not considered in the evaluation process. Effectiveness of this approach is experimentally evaluated for vision-based classification results of CNN considering different image parameters. Two classification models are trained using ResNet and MobileNet V2. The proposed target response approach is implemented on MobileNet V2 classifier. Furthermore, the results of two classifiers are compared using the novel Hit/Miss evaluation approach. The results indicate that Mobile2Net V2 performs better considering brightness as parameter while ResNet performs better considering contrast as varying parameter. Based on the newly introduced POD-related evaluation, different vision-based classification approaches can be clearly distinguished with respect to their ability to predict the correct image as a function of related image parameters. The introduced approach provides a comprehensive interpretation of the quality of a classification model and therefore allows a new evaluation quality.

6.2 Future work

In this thesis, the diagnostic capabilities of different sensor-feature combination is presented. The performance evaluation of these sensors and classification approaches implemented in machine learning and computer vision systems are developed using the POD approach. The evaluation procedure considered a single parameter in the assessment process. As future work, the consideration of two or more parameters concurrently in the evaluation process needs to be developed. This is important because the POD approach demonstrated a strong dependence on the sensor-feature-position combination. In the Machine Learning field, the approach shows a strong dependence on the target response value, while in the computer vision field it showed

a dependence on the image parameter. Therefore an approach considering simultaneously two or more parameters in the evaluation process will help conclude on the general best approach.

Also further research to improve the POD using other fusion strategies needs to be investigated. The introduced POD-based fusion is applied to one example to evaluate the feasibility of the introduced concept. To critically examine the influence on the fusion performance, further experiments have to be undertaken to consider all sensor-feature combinations detecting a fault or not. Using the results of the experiments, individual accuracy and fused accuracy can be compared.

Bibliography

[AASL16] Aldrin, John C. ; Annis, Charles ; Sabbagh, Harold A. ; Lindgren, Eric A.: Best practices for evaluating the capability of nondestructive evaluation (NDE) and structural health monitoring (SHM) techniques for damage characterization. In: *AIP Conference Proceedings* Bd. 1706 AIP Publishing LLC, 2016, pp. 200002

[AB12] Assous, Said ; Boashash, Boualem: Evaluation of the modified S-transform for time-frequency synchrony analysis and source localisation. In: *EURASIP Journal on Advances in Signal Processing* 2012 (2012), Nr. 1, pp. 49

[ABSC11] Al-Badour, F ; Sunar, M ; Cheded, L: Vibration analysis of rotating machinery using time–frequency analysis and wavelet techniques. In: *Mechanical Systems and Signal Processing* 25 (2011), Nr. 6, pp. 2083–2101

[ADS19] Ameyaw, Daniel A. ; Deng, Qi ; Söffker, Dirk: Probability of Detection (POD)-based Metric for Evaluation of Classifiers Used in Driving Behavior Prediction. In: *Proceedings of the Annual Conference of the PHM Society* Bd. 11, 2019

[ADS20a] Ameyaw, Daniel A. ; Deng, Qi ; Söffker, Dirk: A new kind of visualization and interpretation of classifier performance. In: *Information Sciences* (2020). – submitted

[ADS20b] Ameyaw, Daniel A. ; Deng, Qi ; Söffker, Dirk: A New Measure Comparing Machine Learning-based Classification Approaches Applied to Dynamically Changing Situations. In: *IEEE Intelligent Systems* (2020). – submitted

[AML+11] Aldrin, John C. ; Medina, Enrique A. ; Lindgren, Eric A. ; Buynak, Charles F. ; Knopp, Jeremy S.: Case studies for model-assisted probabilistic reliability assessment for structural health monitoring systems. In: *AIP Conference Proceedings* Bd. 1335 American Institute of Physics, 2011, pp. 1589–1596

[Ann17] Annis, Charles: Statistical best-practices for building Probability of Detection (POD) models. In: *R package mh1823, Version* 2 (2017), Nr. 4.4

[ARS18a] Ameyaw, Daniel A. ; Rothe, Sandra ; Söffker, Dirk: Adaptation and implementation of Probability of Detection (POD)-based

fault diagnosis in elastic structures through vibration-based SHM approach. In: *The 9th European Workshop on Structural Health Monitoring (EWSHM), Manchester*, 2018

[ARS18b] Ameyaw, Daniel A. ; Rothe, Sandra ; Söffker, Dirk: Probability of detection (POD)-oriented view to fault diagnosis for reliability assessment of FDI approaches. In: *ASME 2018 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* American Society of Mechanical Engineers Digital Collection, 2018

[ARS18c] Ameyaw, Daniel A. ; Rothe, Sandra ; Söffker, Dirk: Fault diagnosis using Probability of Detection (POD)-based sensor/information fusion for vibration-based analysis of elastic structures. In: *PAMM* 18 (2018), Nr. 1, pp. e201800474

[ARS20] Ameyaw, Daniel A. ; Rothe, Sandra ; Söffker, Dirk: A novel feature-based probability of detection assessment and fusion approach for reliability evaluation of vibration-based diagnosis systems. In: *Structural Health Monitoring* 19 (2020), Nr. 3, pp. 649–660

[AS20] Ameyaw, Daniel A. ; Söffker, Dirk: Fault diagnosis and damage quantication of elastic mechanical systems: State of the art, limitations, and outlook. In: *Mechanical Systems and Signal Processing* (2020). – submitted

[ASSS14] Al-Shrouf, Lou'i ; Saadawia, Mahmud-Sami ; Söffker, Dirk: Improved process monitoring and supervision based on a reliable multi-stage feature-based pattern recognition technique. In: *Information Sciences-Informatics and Computer Science* 259 (2014), pp. 282–294

[BAMS20] Bakhshande, Fateme ; Ameyaw, Daniel A. ; Madan, Neelu ; Söffker, Dirk: Performance Evaluation Approach based on Probability of Detection (POD) for Vision-based Deep Neural Network Classiers. In: *Applied Intelligence* (2020). – submitted

[BJSO14] Britto Jr, Alceu S. ; Sabourin, Robert ; Oliveira, Luiz E.: Dynamic selection of classifiersa comprehensive review. In: *Pattern recognition* 47 (2014), Nr. 11, pp. 3665–3680

[BKD01] Bowyer, Kevin ; Kranenburg, Christine ; Dougherty, Sean: Edge detector evaluation using empirical ROC curves. In: *Computer Vision and Image Understanding* 84 (2001), Nr. 1, pp. 77–103

[BLW+87] Boashash, Boualem ; Lovell, Brian ; Whitehouse, Harper J. [u. a.]: High-resolution time-frequency signal analysis by parametric modelling of the Wigner-ville distribution. In: *1st IASTED International Symposium on Signal Processing and Its Applications* Institution of Engineers, Australia, 1987, pp. 297

[BN+93] Basseville, Michèle ; Nikiforov, Igor V. [u. a.]: *Detection of abrupt changes: theory and application*. Bd. 104. Prentice Hall Englewood Cliffs, 1993

[Bre01] Breiman, Leo: Random forests. In: *Machine learning* 45 (2001), Nr. 1, pp. 5–32

[BS15] Bovsunovsky, Anatoly ; Surace, Cecilia: Non-linearities in the vibrations of elastic structures with a closing crack: a state of the art review. In: *Mechanical Systems and Signal Processing* 62 (2015), pp. 129–148

[Can10] Canal, Mehmet R.: Comparison of wavelet and short time Fourier transform methods in the analysis of EMG signals. In: *Journal of medical systems* 34 (2010), Nr. 1, pp. 91–94

[CB08] Carden, Peter E. ; Brownjohn, James M.: ARMA modelled time-series classification for structural health monitoring of civil infrastructure. In: *Mechanical systems and signal processing* 22 (2008), Nr. 2, pp. 295–314

[CFL03] Chang, Peter C. ; Flatau, Alison ; Liu, SC: Health monitoring of civil infrastructure. In: *Structural health monitoring* 2 (2003), Nr. 3, pp. 257–267

[Che15] Chen, Chi-hau: *Handbook of pattern recognition and computer vision*. World Scientific, 2015

[CI83] Cheng, Russel C. ; Iles, Terence C.: Confidence bands for cumulative distribution functions of continuous random variables. In: *Technometrics* 25 (1983), Nr. 1, pp. 77–86

[CTW94] Chance, Julian ; Tomlinson, Geoffrey R. ; Worden, Keith: A simplified approach to the numerical and experimental modelling of the dynamics of a cracked beam. In: *Proceedings-SPIE the International Society for Optical Engineering* Citeseer, 1994, pp. 778–778

[CV95] Cortes, Corinna ; Vapnik, Vladimir: Support-vector networks. In: *Machine learning* 20 (1995), Nr. 3, pp. 273–297

[DDS+09] Deng, Jia ; Dong, Wei ; Socher, Richard ; Li, Li-Jia ; Li, Kai ; Fei-Fei, Li: Imagenet: A large-scale hierarchical image database. In: *2009 IEEE conference on computer vision and pattern recognition* IEEE, 2009, pp. 248–255

[DFG96] Doebling, Scott W. ; Farrar, Charles R. ; Goodman, Randall S.: Effects of measurement statistics on the detection of damage in the Alamosa Canyon Bridge / Los Alamos National Lab., NM (United States). 1996. – Technical Report

[DFP+98] Doebling, Scott W. ; Farrar, Charles R. ; Prime, Michael B. [u. a.]: A summary review of vibration-based damage identification methods. In: *Shock and vibration digest* 30 (1998), Nr. 2, pp. 91–105

[DFPS96] Doebling, Scott W. ; Farrar, Charles R. ; Prime, Michael B. ; Shevitz, Daniel W.: Damage identification and health monitoring of structural and mechanical systems from changes in their vibration characteristics: a literature review / Los Alamos National Lab., NM (United States). 1996. – Technical Report

[DG06] Davis, Jesse ; Goadrich, Mark: The relationship between Precision-Recall and ROC curves. In: *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240

[DG13] Dai, Xuewu ; Gao, Zhiwei: From model, signal to knowledge: A data-driven perspective of fault detection and diagnosis. In: *IEEE Transactions on Industrial Informatics* 9 (2013), Nr. 4, pp. 2226–2238

[DS18] Deng, Qi ; Söffker, Dirk: Improved driving behaviors prediction based on Fuzzy Logic-Hidden Markov Model (FL-HMM). In: *2018 IEEE Intelligent Vehicles Symposium (IV)* IEEE, 2018, pp. 2003–2008

[DS19] Deng, Qi ; Söffker, Dirk: Classifying Human Behaviors: Improving Training of Conventional Algorithms. In: *2019 IEEE Intelligent Transportation Systems Conference (ITSC)* IEEE, 2019, pp. 1060–1065

[DSJ08] Djurović, Igor ; Sejdić, Ervin ; Jiang, Jin: Frequency-based window width optimization for S-transform. In: *AEU-International Journal of Electronics and Communications* 62 (2008), Nr. 4, pp. 245–250

[DSP+16] Dos Santos, Fábio Luis M. ; Peeters, Bart ; ; Van Der Auweraer, Herman ; Góes, Roland ; Desmet, Wim: Vibration-based damage detection for a composite helicopter main rotor blade. In: *Case Studies in Mechanical Systems and Signal Processing* 3 (2016), pp. 22–27

[DWS18] Deng, Qi ; Wang, Jiao ; Söffker, Dirk: Prediction of human driver behaviors based on an improved HMM approach. In: *2018 IEEE Intelligent Vehicles Symposium (IV)* IEEE, 2018, pp. 2066–2071

[Faw06] Fawcett, Tom: An introduction to ROC analysis. In: *Pattern recognition letters* 27 (2006), Nr. 8, pp. 861–874

[FDCS96] Farrar, Charles R. ; Doebling, Scott W. ; Cornwell, Phillip J. ; Straser, Erik G.: Variability of modal parameters measured on the Alamosa Canyon Bridge / Los Alamos National Lab., NM (United States). 1996. – Technical Report

[FDDN99] Farrar, Charles R. ; Duffey, Thomas A. ; Doebling, Scott W. ; Nix, David A.: A statistical pattern recognition paradigm for vibration-based structural health monitoring. In: *Structural Health Monitoring* 2000 (1999), pp. 764–773

[FDN01] Farrar, Charles R. ; Doebling, Scott W. ; Nix, David A.: Vibration–based structural damage identification. In: *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 359 (2001), Nr. 1778, pp. 131–149

[FK09] Fritzen, Claus-Peter ; Kraemer, Peter: Self-diagnosis of smart structures based on dynamical properties. In: *Mechanical Systems and Signal Processing* 23 (2009), Nr. 6, pp. 1830–1845

[Fox92] Fox, Colin H.: The location of defects in structures-A comparison of the use of natural frequency and mode shape data. In: *10th International modal analysis conference* Bd. 1, 1992, pp. 522–528

[Fri05] Fritzen, Claus P.: Vibration-based structural health monitoring–concepts and applications. In: *Key Engineering Materials* Bd. 293 Trans Tech Publ, 2005, pp. 3–20

[FS09] Fassois, Spilios D. ; Sakellariou, John S.: Statistical time series methods for SHM. In: *Encyclopedia of structural health monitoring* (2009)

[Fug09] Fugal, Lee D.: Preview of Wavelets, Wavelet Filters and Wavelet Transforms. In: *Conceptual Wavelets In Digital Signal Processing* (2009), pp. 1–30

[FW07] Farrar, Charles R. ; Worden, Keith: An introduction to structural health monitoring. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 365 (2007), Nr. 1851, pp. 303–315

[GA10] Gandossi, Luca ; Annis, Charles: Probability of detection curves: Statistical best-practices. In: *ENIQ report* 41 (2010)

[GC09] Gul, Mustafa ; Catbas, Necati F.: Statistical pattern recognition for Structural Health Monitoring using time series modeling: Theory and experimental verifications. In: *Mechanical Systems and Signal Processing* 23 (2009), Nr. 7, pp. 2192–2204

[Geo07] Georgiou, George: PoD curves, their derivation, applications and limitations. In: *Insight-Non-Destructive Testing and Condition Monitoring* 49 (2007), Nr. 7, pp. 409–414

[HHB09] Harding, Cayt ; Hugo, Geoff ; Bowles, Susan: Application of model-assisted POD using a transfer function approach. In: *AIP Conference Proceedings* Bd. 1096 American Institute of Physics, 2009, pp. 1792–1799

[HS93] Huang, Yea S. ; Suen, Ching Y.: The behavior-knowledge space method for combination of multiple classifiers. In: *IEEE computer society conference on computer vision and pattern recognition* Institute of Electrical Engineers Inc (IEEE), 1993, pp. 347–347

[HSL+98] Huang, Norden E. ; Shen, Zheng ; Long, Steven R. ; Wu, Manli C. ; Shih, Hsing H. ; Zheng, Quanan ; Yen, Nai-Chyuan ; Tung, Chi C. ; Liu, Henry H.: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. In: *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences* 454 (1998), Nr. 1971, pp. 903–995

[HZRS15] He, Kaiming ; Zhang, Xiangyu ; Ren, Shaoqing ; Sun, Jian: Spatial pyramid pooling in deep convolutional networks for visual recognition. In: *IEEE transactions on pattern analysis and machine intelligence* 37 (2015), Nr. 9, pp. 1904–1916

[HZRS16] He, Kaiming ; Zhang, Xiangyu ; Ren, Shaoqing ; Sun, Jian: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778

[HZY07] Hongbin, Li ; Zhao, Qing ; Yang, Zhenyu: Reliability modeling of fault tolerant control systems. In: *International Journal of Applied Mathematics and Computer Science* 17 (2007), Nr. 4, pp. 491–504

[IB97] Isermann, Rolf ; Balle, Peter: Trends in the application of model-based fault detection and diagnosis of technical processes. In: *Control engineering practice* 5 (1997), Nr. 5, pp. 709–719

[Ise06] Isermann, Rolf: *Fault-diagnosis systems: an introduction from fault detection to fault tolerance*. Springer Science & Business Media, 2006

[JVW10] Jayaswal, Pratesh ; Verma, Sunita N. ; Wadhwani, Arun K.: Application of ANN, fuzzy logic and wavelet transform in machine fault diagnosis using vibration signal analysis. In: *Journal of Quality in Maintenance Engineering* 16 (2010), Nr. 2, pp. 190–213

[KALA07] Knopp, Jeremy S. ; Aldrin, John C. ; Lindgren, Eric A. ; Annis, Charles: Investigation of a Model-Assisted Approach to Probability of Detection Evaluation. In: *AIP Conference Proceedings* Bd. 894 American Institute of Physics, 2007, pp. 1775–1782

[KBD01] Kuncheva, Ludmila I. ; Bezdek, James C. ; Duin, Robert P.: Decision templates for multiple classifier fusion: an experimental comparison. In: *Pattern recognition* 34 (2001), Nr. 2, pp. 299–314

[KF10] Kopsaftopoulos, Fotis P. ; Fassois, Spilios D.: Vibration based health monitoring for a lightweight truss structure: experimental assessment of several statistical time series methods. In: *Mechanical Systems and Signal Processing* 24 (2010), Nr. 7, pp. 1977–1997

[KF13] Kopsaftopoulos, Fotis P. ; Fassois, Spilios D.: A functional model based statistical time series method for vibration based damage detection, localization, and magnitude estimation. In: *Mechanical Systems and Signal Processing* 39 (2013), Nr. 1-2, pp. 143–161

[KN05] Ko, Jinming ; Ni, Yi-Qing: Technology developments in structural health monitoring of large-scale bridges. In: *Engineering structures* 27 (2005), Nr. 12, pp. 1715–1725

[KNN+05] Kutner, Michael H. ; Nachtsheim, Christopher J. ; Neter, John ; Li, William [u. a.]: *Applied linear statistical models*. Bd. 5. McGraw-Hill Irwin New York, 2005

[LAGS+13] Lorente, Delia ; Aleixos, Nuria ; Gómez-Sanchis, Juan ; Cubero, Sergio ; Blasco, Jose: Selection of optimal wavelength features for decay detection in citrus fruit using the ROC curve and neural networks. In: *Food and Bioprocess Technology* 6 (2013), Nr. 2, pp. 530–541

[Liu13] Liu, WY: The vibration analysis of wind turbine blade–cabin–tower coupling system. In: *Engineering Structures* 56 (2013), pp. 954–957

[LJL+19] Liu, Bo ; Jing, Liping ; Li, Jia ; Yu, Jian ; Gittens, Alex ; Mahoney, Michael W.: Group collaborative representation for image

set classification. In: *International Journal of Computer Vision* 127 (2019), Nr. 2, pp. 181–206

[LMSM+16] Lazarova-Molnar, Sanja ; Shaker, Hamid R. ; Mohamed, Nader [u. a.]: Fault detection and diagnosis for smart buildings: State of the art, trends and challenges. In: *2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC)* IEEE, 2016, pp. 1–7

[LOW+20] Liu, Li ; Ouyang, Wanli ; Wang, Xiaogang ; Fieguth, Paul ; Chen, Jie ; Liu, Xinwang ; Pietikäinen, Matti: Deep learning for generic object detection: A survey. In: *International Journal of Computer Vision* 128 (2020), Nr. 2, pp. 261–318

[LW98] Liew, Kim M. ; Wang, Quan: Application of wavelet theory for crack identification in structures. In: *Journal of engineering mechanics* 124 (1998), Nr. 2, pp. 152–157

[Mal99] Mallat, Stéphane: *A wavelet tour of signal processing*. Elsevier, 1999

[Mar12] Marwala, Tshilidzi: *Condition monitoring using computational intelligence methods: applications in mechanical and electrical systems*. Springer Science & Business Media, 2012

[May91] Mayes, Randy: Error localization using mode shapes: An application to a two link robot arm / Sandia National Labs., Albuquerque, NM (United States). 1991. – Technical Report

[MDK+97] Martin, Alvin ; Doddington, George ; Kamm, Terri ; Ordowski, Mark ; Przybocki, Mark: The DET curve in assessment of detection task performance / National Inst of Standards and Technology Gaithersburg MD USA. 1997. – Technical Report

[ME15] Mousavi, Reza ; Eftekhari, Mahdi: A new ensemble learning methodology based on hybridization of classifier ensemble selection approaches. In: *Applied Soft Computing* 37 (2015), pp. 652–666

[MHA09] MIL-HDBK-1823A: Department of Defense Handbook: Nondestructive Evaluation System Reliability Assessment. (2009)

[MJ11] Ma, Jianping ; Jiang, Jin: Applications of fault detection and diagnosis methods in nuclear power plants: A review. In: *Progress in nuclear energy* 53 (2011), Nr. 3, pp. 255–266

[MJB+11] Müller, Ingolf ; Janapati, Vishnuvardhan ; Banerjee, Sourav ; Lonkar, K. ; Roy, Surajit ; Chang, Fu-kuo: On the performance

quantification of active sensing SHM systems using model-assisted POD methods. In: *Structural Health Monitoring* 2 (2011), pp. 2417–2428

[MMBC13] Mukhopadhyay, Anirban ; Maulik, Ujjwal ; Bandyopadhyay, Sanghamitra ; Coello, Carlos Artemio C.: A survey of multiobjective evolutionary algorithms for data mining: Part I. In: *IEEE Transactions on Evolutionary Computation* 18 (2013), Nr. 1, pp. 4–19

[MQLDM18] Moriot, Jérémy ; Quaegebeur, Nicolas ; Le Duff, Alain ; Masson, Patrice: A model-based approach for statistical assessment of detection and localization performance of guided wave–based imaging techniques. In: *Structural Health Monitoring* 17 (2018), Nr. 6, pp. 1460–1472

[MS08] Mortazavi, SH ; Shahrtash, SM: Comparing denoising performance of DWT, WPT, SWT and DT-CWT for partial discharge signals. In: *2008 43rd International Universities Power Engineering Conference* IEEE, 2008, pp. 1–6

[MU10] Mendrok, Krzysztof ; Uhl, Tadeusz: The application of modal filters for damage detection. In: *Smart structures and systems* 6 (2010), Nr. 2, pp. 115–133

[NW72] Nelder, John A. ; Wedderburn, Robert W.: Generalized linear models. In: *Journal of the Royal Statistical Society: Series A (General)* 135 (1972), Nr. 3, pp. 370–384

[OPB12] Oshiro, Thais M. ; Perez, Pedro S. ; Baranauskas, José A.: How many trees in a random forest? In: *International workshop on machine learning and data mining in pattern recognition* Springer, 2012, pp. 154–168

[PBS91] Pandey, AK ; Biswas, M ; Samman, MM: Damage detection from changes in curvature mode shapes. In: *Journal of sound and vibration* 145 (1991), Nr. 2, pp. 321–332

[PGO14] Perez, Marco A. ; Gil, Lluis ; Oller, Sergio: Impact damage identification in composite laminates using vibration testing. In: *Composite Structures* 108 (2014), pp. 267–276

[PH] Pettit, Donald E. ; Hoeppner, David W.: Fatigue flaw growth and NDI evaluation for preventing through cracks in spacecraft tankage structures. In: *NAS 9-11722, LR 25387*

[PJK+16] Poon, Jason ; Jain, Palak ; Konstantakopoulos, Ioannis C. ; Spanos, Costas ; Panda, Sanjib K. ; Sanders, Seth R.: Model-based fault detection and identification for switching power converters. In: *IEEE Transactions on Power Electronics* 32 (2016), Nr. 2, pp. 1419–1430

[PM09] Polimeno, U ; Meo, M: Detecting barely visible impact damage detection on aircraft composites structures. In: *Composite structures* 91 (2009), Nr. 4, pp. 398–402

[Qin12] Qin, S J.: Survey on data-driven industrial process monitoring and diagnosis. In: *Annual reviews in control* 36 (2012), Nr. 2, pp. 220–234

[Rab89] Rabiner, Lawrence R.: A tutorial on hidden Markov models and selected applications in speech recognition. In: *Proceedings of the IEEE* 77 (1989), Nr. 2, pp. 257–286

[RG00] Ruta, Dymitr ; Gabrys, Bogdan: An overview of classifier fusion methods. In: *Computing and Information systems* 7 (2000), Nr. 1, pp. 1–10

[RJS15] Rother, Astrid ; Jelali, Mohieddine ; Söffker, Dirk: A brief review and a first application of time-frequency-based analysis methods for monitoring of strip rolling mills. In: *Journal of Process Control* 35 (2015), pp. 65–79

[RKP11] Radzieński, Maciej ; Krawczuk, Marek ; Palacz, Magdalena: Improvement of damage detection methods based on experimental modal parameters. In: *Mechanical Systems and Signal Processing* 25 (2011), Nr. 6, pp. 2169–2190

[RKS19] Rothe, Sandra ; Kudszus, Bastian ; Söffker, Dirk: Does classifier fusion improve the overall performance? Numerical analysis of data and fusion method characteristics influencing classifier fusion performance. In: *Entropy* 21 (2019), Nr. 9, pp. 866

[Rot19] Rothe, Sandra: *Reliable information fusion methods for condition monitoring*, University of Duisburg-Essen, Diss., 2019

[RS16] Rothe, Sandra ; Söffker, Dirk: Comparison of different information fusion methods using ensemble selection considering benchmark data. In: *2016 19th International Conference on Information Fusion (FUSION)* IEEE, 2016, pp. 73–78

[SFHW01] Sohn, Hoon ; Farrar, Charles R. ; Hunter, Norman F. ; Worden, Keith: Structural health monitoring using statistical pattern recognition techniques. In: *J. Dyn. Sys., Meas., Control* 123 (2001), Nr. 4, pp. 706–711

[SHZ+14] Sun, Hailiang ; He, Zhengjia ; Zi, Yanyang ; Yuan, Jing ; Wang, Xiaodong ; Chen, Jinglong ; He, Shuilong: Multiwavelet transform and its applications in mechanical fault diagnosis–a review. In: *Mechanical Systems and Signal Processing* 43 (2014), Nr. 1-2, pp. 1–24

[SHZ+18] Sandler, Mark ; Howard, Andrew ; Zhu, Menglong ; Zhmoginov, Andrey ; Chen, Liang-Chieh: Mobilenetv2: Inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520

[SK96] Stubbs, Norris ; Kim, Jeong-Tae: Damage localization in structures without baseline modal parameters. In: *AIAA journal* 34 (1996), Nr. 8, pp. 1644–1649

[SKGDD15] Schubert Kabban, Christine M. ; Greenwell, Brandon M. ; DeSimio, Martin P. ; Derriso, Mark M.: The probability of detection for structural health monitoring systems: Repeated measures data. In: *Structural Health Monitoring* 14 (2015), Nr. 3, pp. 252–264

[SKT92] Stubbs, Norris ; Kim, JT ; Topole, K: An efficient and robust algorithm for damage localization in offshore platforms. In: *Proceedings of the ASCE 10th structures congress* Bd. 1, 1992, pp. 543–546

[SL00] Suen, Ching Y. ; Lam, Louisa: Multiple classifier combination methodologies for different output levels. In: *International workshop on multiple classifier systems* Springer, 2000, pp. 52–66

[SL09] Sokolova, Marina ; Lapalme, Guy: A systematic analysis of performance measures for classification tasks. In: *Information processing & management* 45 (2009), Nr. 4, pp. 427–437

[SSB10] Sakthivel, NR ; Sugumaran, V ; Babudevasenapati, Sjeswa: Vibration based fault diagnosis of monoblock centrifugal pump using decision tree. In: *Expert Systems with Applications* 37 (2010), Nr. 6, pp. 4040–4049

[Sta10] Stanković, Srdjan: Time-frequency analysis and its application in digital watermarking. In: *EURASIP Journal on Advances in Signal Processing* 2010 (2010), Nr. 1, pp. 579295

[Sus04] Susmaga, Robert: Confusion matrix visualization. In: *Intelligent Information Processing and Web Mining*. Springer, 2004, pp. 107–116

[SUS13] Stepinski, Tadeusz ; Uhl, Tadeusz ; Staszewski, Wieslaw: *Advanced structural damage detection: from theory to engineering applications*. John Wiley & Sons, 2013

[SWWS16] Soeffker, Dirk ; Wei, Chunsheng ; Wolff, Stefan ; Saadawia, Mahmud-Sami: Detection of rotor cracks: comparison of an old model-based approach with a new signal-based approach. In: *Nonlinear Dynamics* 83 (2016), Nr. 3, pp. 1153–1170

[TBF+09] Thompson, R B. ; Brasche, Lisa H. ; Forsyth, David ; Lindgren, Eric ; Swindell, Paul ; Winfree, W: Recent advances in model-assisted probability of detection. (2009)

[TCO08] Trendafilova, I ; Cartmell, Matthew P. ; Ostachowicz, Wieslaw: Vibration-based damage detection in an aircraft wing scaled model using principal component analysis and pattern recognition. In: *Journal of Sound and Vibration* 313 (2008), Nr. 3-5, pp. 560–566

[TJGD08] Tulyakov, Sergey ; Jaeger, Stefan ; Govindaraju, Venu ; Doermann, David: Review of classifier combination methods. In: *Machine learning in document analysis and recognition*. Springer, 2008, pp. 361–386

[TS95a] Topole, Klaus G. ; Stubbs, Norris: Non-destructive damage evaluation of a structure from limited modal parameters. In: *Earthquake engineering & structural dynamics* 24 (1995), Nr. 11, pp. 1427–1436

[TS95b] Topole, Klaus G. ; Stubbs, Norris: Nondestructive damage evaluation in complex structures from a minimum of modal parameters. In: *International Journal of Analytical and Modal Analysis* 1 (1995), pp. 95–103

[TSK+18] Tan, Chuanqi ; Sun, Fuchun ; Kong, Tao ; Zhang, Wenchang ; Yang, Chao ; Liu, Chunfang: A survey on deep transfer learning. In: *International conference on artificial neural networks* Springer, 2018, pp. 270–279

[WD99] Wang, Quan ; Deng, Xiaomin: Damage detection with spatial wavelets. In: *International journal of solids and structures* 36 (1999), Nr. 23, pp. 3443–3468

[Wes86] West, Walter M.: Illustration of the use of modal assurance criterion to detect structural changes in an orbiter test specimen. In: *4th*

International modal analysis conference, Los Angeles Bd. 1, 1986, pp. 1–6

[WFY15] Wang, Zilei ; Feng, Jiashi ; Yan, Shuicheng: Collaborative linear coding for robust image classification. In: *International Journal of Computer Vision* 114 (2015), Nr. 2-3, pp. 322–333

[WLL+13] Wang, Xuelei ; Li, Qingmin ; Li, Chengrong ; Yang, Rui ; Su, Qi: Reliability assessment of the fault diagnosis methodologies for transformers and a new diagnostic scheme based on fault info integration. In: *IEEE Transactions on Dielectrics and Electrical Insulation* 20 (2013), Nr. 6, pp. 2292–2298

[WY07] Widodo, Achmad ; Yang, Bo-Suk: Support vector machine in machine condition monitoring and fault diagnosis. In: *Mechanical systems and signal processing* 21 (2007), Nr. 6, pp. 2560–2574

[XKS92] Xu, Lei ; Krzyzak, Adam ; Suen, Ching Y.: Methods of combining multiple classifiers and their applications to handwriting recognition. In: *IEEE Transactions on System, Man, and Cybernetics* 22 (1992), Nr. 3, pp. 418–435

[YS13] Yang, Bin ; Sun, Dongbai: Testing, inspecting and monitoring technologies for wind turbine blades: A survey. In: *Renewable and Sustainable Energy Reviews* 22 (2013), pp. 515–526

[ZXL+16] Zou, Quan ; Xie, Sifa ; Lin, Ziyu ; Wu, Meihong ; Ju, Ying: Finding the best classification threshold in imbalanced classification. In: *Big Data Research* 5 (2016), pp. 2–8

[ZZXW19] Zhao, Zhong-Qiu ; Zheng, Peng ; Xu, Shou-Tao ; Wu, Xindong: Object detection with deep learning: A review. In: *IEEE transactions on neural networks and learning systems* 30 (2019), Nr. 11, pp. 3212–3232

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

Publications

[ARS20] Ameyaw, D. A.; Rothe, S.; Söffker, D.: A novel feature-based probability of detection assessment and fusion approach for reliability evaluation of vibration-based diagnosis systems. *Structural Health Monitoring* 19, no.3 (2020), pp. 649-660, https://doi.org/10.1177/1475921719856274

[ADS20a] Ameyaw, D. A.; Deng, Q.; Söffker, D.: A new kind of visualization and interpretation of classifier performance. In: *Information Sciences*, submitted

[ADS20b] Ameyaw, D. A.; Deng, Q.; Söffker, D.: A new measure comparing machine learning-based classification approaches applied to dynamically changing situations. In: *IEEE Intelligent Systems*, submitted

[BAMS20] Bakhshande, F.; Ameyaw, D. A.; Madan, N.; Söffker, D.: Performance Evaluation Approach based on Probability of Detection (POD) for Vision-based Deep Neural Network Classifiers. In: *Applied Intelligence*, submitted

[ADS19] Ameyaw, D. A.; Deng, Q.; Söffker, D.: Probability of Detection (POD)-based metric for evaluation of Classifiers used in Driving Behavior Prediction. In: *Proceedings of the Annual Conference of the PHM Society*, Sep. 22 2019, vol. 11, no. 1

[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. In: *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 and Computers and Information in Engineering Conference. 30th Conference on Mechanical Vibration and Noise, Quebec, Canada*, 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. In: *PAMM 18*, (2018): no. 1, pp. e201800474.

As part of the responsibilities, doctoral candidates exercise effective supervisory role under the guidance of the Chair head. With regards to research projects at the Chair of Dynamics and Control, the following student theses have been successfully supervised by Daniel Adofo Ameyaw and Univ.-Prof. Dr.-Ing. Dirk Söffker.
The research projects and the results from these theses are not integrated in this thesis.

[El15] El-Bahrawy, Y., Anwendung von statischer Mustererkennung zur datenbasierten Schadenerkennung in Wälzlagern anhand von Benchmarkdaten, Master Thesis, August 2017

[Muh19] Muhammad, M. K., Construction of elastic multi-storey structure, Bachelor Thesis, May 2019

DOI: 10.17185/duepublico/72779
URN: urn:nbn:de:hbz:464-20200914-151433-7