Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

How to evaluate classifier performance in the presence of additional effects: A new POD-based approach allowing certification of machine learning approaches

Daniel Adofo Ameyaw [*], Qi Deng, Dirk Söffker

Chair of Dynamics and Control, University of Duisburg–Essen, Lotharstraße 1-21, 47057, Duisburg, Germany

ARTICLE INFO

Keywords:
Performance evaluation
Classifier
Probability of Detection
Human behavior prediction

ABSTRACT

Classifiers are useful and well-known machine learning algorithms allowing classifications. A classifier may be suited for a specific task depending on the application and datasets. To select an approach for a task, performance evaluation may be imperative. Existing approaches like the receiver operating characteristic and precision–recall curves are popular in evaluating classifier performance, however both measures do not directly address the influence of additional and possibly unknown (process) parameters on the classification results. In this contribution, this limitation is discussed and addressed by adapting the Probability of Detection (POD) measure. The POD is a probabilistic method to quantify the reliability of a diagnostic procedure taking into account statistical variability of sensor and measurements properties. In this contribution the POD approach is adapted and extended. The introduced approach is implemented on driving behavior prediction data serving as illustrative example. Based on the introduced POD-related evaluation, different classifiers can be clearly distinguished with respect to their ability to predict the correct intended driver behavior as a function of remaining time (here assumed as process parameter) before the event itself. The introduced approach provides a new diagnostic and comprehensive interpretation of the quality of a classification model.

1. Introduction

Machine Learning (ML) have been developed and used over decades in many applications. Advancements in processing capabilities of computers has improved implementation time. Institutions and researchers are using ML to analyze data patterns, detect fake news online, predict consumer behavior, detect frauds in financial transactions, clinical trials, predict behavioral tendencies, and statistical analysis to literary works (Ali, Miah, Haque, Rahman, & Islam, 2021; Carcillo et al., 2019; González-Carrasco, Jiménez-Márquez, López-Cuadrado, & Ruiz-Mezcua, 2019; Khan, Khondaker, Afroz, Uddin, & Iqbal, 2021; Lin et al., 2020; Peng & Hengartner, 2002). Useful algorithms to assign class labels at specific data points are denoted as classifiers. Performance assessment relative to the core functionality of classifiers is required to determine the reliability of related approaches. These performance evaluation is useful and typically applied in selecting a classifier and/or comparing different approaches for a specific task. The Receiver Operating Characteristic (ROC) and Precision–Recall (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. While the ROC curve uses the ratio of Detection Rate (DR) to False Alarm Rate (FAR), the PR curve utilize the ratio of precision to recall therefore allowing direct comparison of diagnostic tests for a specific application. Despite the popularity of these evaluation processes, the effect of process parameters on classification results is not directly addressed (Ameyaw, Deng, & Söffker, 2019). Process parameters should not be misunderstood as learning/tuning hyperparameters. Process parameters are defined according to Ameyaw et al. (2019), DOD (2009) as:

i. Parameters of the process/task that has influence on the classification result.
ii Parameter with which the recognizability is qualitatively or quantitatively influenced.

The conventional use of ROC/PR measures is based on the overlap probabilities for signal and measurement noise (see Fig. 1).

The area under ROC (AUC) is a widely used performance measure. The one-pass AUC optimization has been suggested for going through the training data only once without having to store the entire training dataset (Gao, Wang, Jin, Zhu, & Zhou, 2016). However, authors criticize the use of AUC as it is dominated by high FAR points (Bowyer, Kranenburg, & Dougherty, 2001; Lorente, Aleixos, Gómez-Sanchis, Cubero, & Blasco, 2013). The PR curve has been proposed as more effective in dealing with highly skewed datasets (Cook

[*] Corresponding author.
E-mail addresses: daniel.adofo-ameyaw@uni-due.de (D.A. Ameyaw), qi.deng@uni-due.de (Q. Deng), soeffker@uni-due.de (D. Söffker).

https://doi.org/10.1016/j.mlwa.2021.100220
Received 17 September 2021; Received in revised form 11 November 2021; Accepted 12 November 2021
Available online 26 November 2021

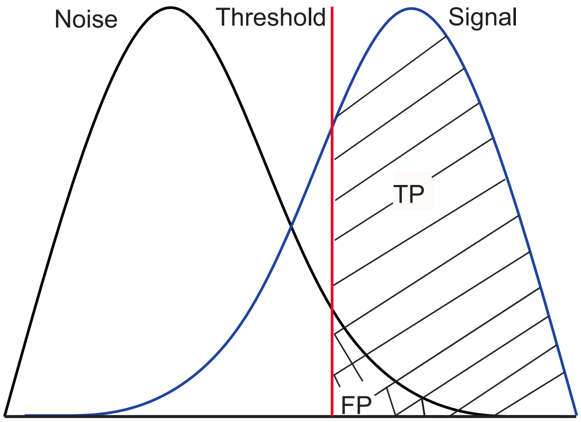


Fig. 1. Overlapping probability densities of signal and noise.

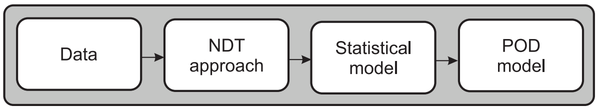


Fig. 2. Existing POD approach.

& Ramadas, 2020; Ozenne, Subtil, & Maucort-Boulch, 2015). However in several engineering application fields classification tasks are defined not only by the main relevant data. Typically used for training and testing are input-/output relations. Often further (non-modeled) effects affect the classification results. This fundamental concern of investigating the effect of process parameters on classification results has received little attention (Ameyaw et al., 2019). This contribution attempts to address this concern using the Probability of Detection (POD) reliability measure and a newly developed visual representation. The POD will be adapted, modified, and implemented in evaluating classifier performance by relating detectability quantitatively to process parameter.

Probability of detection measure is a certification standard that has been implemented in the Nondestructive testing (NDT), Structural health monitoring (SHM), and material testing fields (Georgiou, 2007). The POD is a probabilistic method to quantify the reliability of an NDT, SHM or material testing procedure taking into account statistical variability of sensor and measurements properties (Ameyaw, Rothe, & Söffker, 2020; DOD, 2009). Existing evaluation procedure is illustrated in Fig. 2.

The NDT field utilize the so-called POD curve. The POD curve is constructed by plotting the accumulation of flaws detected against the varying parameter or produce a response over a specified threshold (Georgiou, 2007; Ginzel, 2006). The curve relates probability of detecting a flaw with severity. Size is usually used as a proxy for severity. The POD generator presented by Volker, Dijkstra, Terpstra, Heerings, and Lont (2004), allows assessment and optimization of an inspection program for in-service components. The statistical assessments of a measurement procedure is time-consuming and costly considering several samples have to be verified and compared using destructive methods. This 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 (Harding, Hugo, & Bowles, 2009; Knopp, Aldrin, Lindgren, & Annis, 2007; Thompson, Brasche, Forsyth, Lindgren, Swindell, & Winfree, 2009).

Predicting drivers intention is useful in ensuring driving safety in autonomous and/or assisted driving. Though many approaches exist (Qiu, Rachedi, Sallak, & Vanderhaegen, 2017), contemporary applications use mainly ML approaches (Kukkala, Tunnell, Pasricha, & Bradley, 2018; Theissler, Pérez-Velázquez, Kettelgerdes, & Elger, 2021). An important tool used in these predictions are classifiers. A key idea is to establish models by learning from the given driving behaviors and subsequently predict the decisions and behaviors. Research in this field is concerned with new methods to realize and improve driving behavior prediction. However the need to ensure the integrity of sensors, systems, data, and information systems (like classifiers) is growing (Rong, Teixeira, & Soares, 2020). Accordingly there is a need to ensure suitable related reliability requirements. In this contribution the POD approach is adapted and extended to allow the evaluation of classifier performance as illustrated in Fig. 3. As example, data from driving simulator are used to demonstrate the proposed approach.

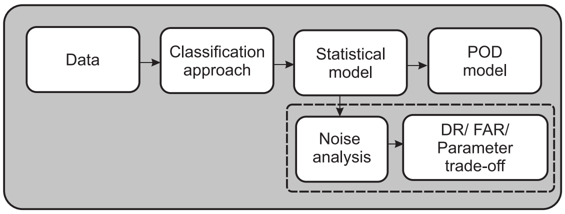


Fig. 3. Modified POD approach.

In a previous publication by Ameyaw et al. (2019), the authors introduced POD evaluation of classifiers and applied this to a first illustrative example. In addition to the previous publication here:

i. The approach is fully developed and compared to classical approaches (F-measure and ROC).
ii Eight classifiers (compared to three) comprising conventional and modified classifiers are examined.
iii A novel noise analysis procedure is introduced fully illustrating the visualization between the DR, FAR, and the process parameter.

The article is organized as follows: in Section 2 the classifiers used as example are briefly introduced, followed by the theoretical derivation of the developed POD reliability measure in Section 3. Experimental validation of the proposed approach and a novel evaluation procedure incorporating process parameter are presented in Section 4. Results, discussions, and comparison between different classifier performance using the developed approach are given in Section 5. A conclusion in Section 6 finalizes the contribution.

2. Binary classifiers

A very brief presentation on the classifiers used for illustration in this paper is presented. Though aspects are known to the ML community, a strategy to improve conventional classifiers presented in Deng and Söffker (2019) is repeated because it is later evaluated to ascertain the effect on the POD. Deng and Söffker (2019) 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 conventional algorithms. In this contribution, eight classifiers; conventional/improved SVM, HMM, ANN, and RF are used. These classifiers are selected to illustrate the proposed approach and therefore not exhaustive. The classifiers prediction abilities will later be evaluated using the POD method.

2.1. Conventional/improved Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning method and a widely applied classification technique (Cortes & Vapnik, 1995). In this contribution, the various classes in SVM refer to different driving behaviors. The driver prediction model in Deng and Söffker (2019) is a multiclass problem (lane change to right S_1, lane keeping S_2, and lane change to left S_3), therefore one-against-one approach is utilized to establish a multiclass SVM 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 (Deng & Söffker, 2018, 2019) applied to define features and influence the prediction performance of SVM.

2.2. Conventional/improved Hidden Markov Models

Hidden Markov Model (HMM) (Rabiner, 1989) describes the relationship between two stochastic processes consisting of a set of unobserved (hidden) states and a set of observable symbols. Due to dynamical changes, the process considered moves from one state to another generating hidden states sequence and observation sequence. Based on a given HMM, the most probable hidden states sequence can be calculated by analyzing the observation sequence. In this work, three different driving behaviors (S_1, S_2, S_3) are modeled as hidden states for the HMM. The related training method of conventional HMM is referred to in Deng, Wang, and Söffker (2018). Here an improved HMM is trained with an optimal prefilter to be designed (Deng & Söffker, 2019). Hidden Markov Model is a suitable algorithm due to its ability to handle time series data and state transition descriptions. Based on observations (training), the HMM approach can calculate the most possible driving behaviors using observed sequences. To improve the prediction performance of the model, a prefilter is proposed to quantize the collected signals into observed sequences with specific features. Here optimality is defined as the optimal segments describing a quantized prefilter mapping the vehicle's environment to quantized states.

2.3. Conventional/improved Artificial Neural Network

Artificial Neural Network (ANN) is a computational model that imitates biological neural network. These models learn to perform tasks without explicit task-specific rules. However, the output node of ANN is a decimal value. To determine final results, usually cut-off thresholds are used to distinguish the decimal values into class labels. Therefore, cut-off thresholds are considered as the tuning parameter of ANN. Related parameters of a conventional/modified ANN (Deng & Söffker, 2019) are used.

2.4. Conventional/improved Random Forest

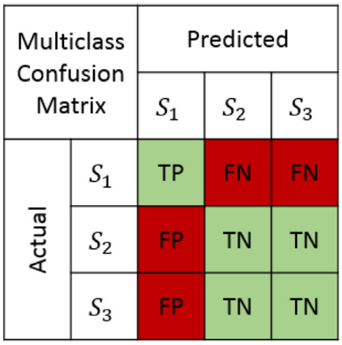
Random Forest (RF) is an extension of decision tree method and it is used to solve classification or regression problems (Breiman, 2001). The RF algorithm contains a set of randomized decision trees that are independent of each other. The total number of decision trees N_{Tree} are unknown and should be defined before the training process. In Oshiro, Perez, and Baranauskas (2012), 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 Deng and Söffker (2019). Similarly, raw data and default N_{Tree} are used to train a conventional RF.

The evaluation are based on True Positive (TP) and False Negative (FN) values calculated using 4.

The DR and FAR values can be defined by Mukhopadhyay, Maulik, Bandyopadhyay, and Coello (2013)

$$DR = \frac{TP}{TP + FN}, and \tag{1}$$

$$FAR = \frac{FP}{TN + FP}. \tag{2}$$



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

Fig. 4. S_1 multiclass confusion matrix.

3. Probability of Detection-based assessment of classifiers

Probability of Detection is a reliability certification tool (Georgiou, 2007). 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 (DOD, 2009; Georgiou, 2007)

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

The target-response approach is adapted in this work because the data to be analyzed relates a changing parameter to its performance response quantitatively. Therefore the approach is adapted and implemented here in evaluating lane changing prediction capabilities of different ML algorithms.

3.1. Target-response approach to POD

The Target-response approach is used when there exist a relationship between a dependent function and an independent variable (DOD, 2009). In the derivation of the POD curve, a predictive modeling technique is required. One such method is regression analysis of the data gathered (Annis, 2020; DOD, 2009; Gandossi & Annis, 2010). 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 (Kutner, Nachtsheim, Neter, Li, et al., 2005). The regression equation for a line of best fit to a given dataset is given by

$$y = b + mx, \tag{3}$$

where m is the slope and b the intercept. The Wald method is used to construct confidence bounds (Kutner et al., 2005). Here the 95% Wald confidence bounds on y is constructed by

$$y_{a=0.95} = y + 1.645\tau_y, \tag{4}$$

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 (Annis, 2020; DOD, 2009). 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

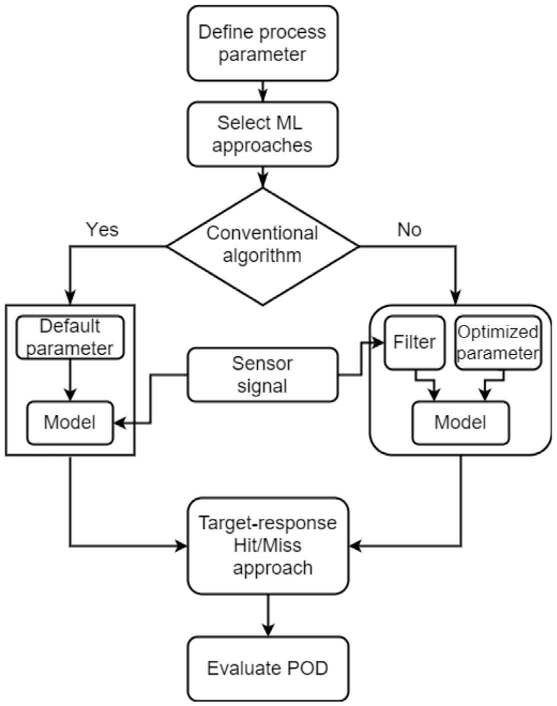


Fig. 5. Classifier POD model building process.

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

$$z_i = \frac{(y_i - (b + mx_i))}{\tau} \tag{5}$$

and

$$f_i = \prod_{i=1}^{n} \frac{1}{2\pi} e^{-\frac{1}{2}(z_i)^2} \tag{6}$$

the information matrix I can be computed as

$$I_{ij} = -E(\frac{\partial}{\partial\Theta_i \partial\Theta_j} log(f)). \tag{7}$$

The inverse of the information matrix yields

$$I^{-1} = \begin{bmatrix} \sigma_b^2 & \sigma_b\sigma_m & \sigma_b\sigma_\tau \\ \sigma_m\sigma_b & \sigma_m^2 & \sigma_m\sigma_\tau \\ \sigma_\tau\sigma_b & \sigma_\tau\sigma_m & \sigma_\tau^2 \end{bmatrix}. \tag{8}$$

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

$$\Phi(\mu, \sigma) = \frac{1}{2}\left[1 + erf\,\frac{x-\mu}{\sqrt{2}\sigma}\right]. \tag{9}$$

The POD as function of target a is derived as

$$POD(a) = \Phi\left[\frac{a-\mu}{\sigma}\right]. \tag{10}$$

Using Eq. (10), the POD-curve can be set up for varying parameters. In this article, the curve is generated for a dynamic system with the varying parameter time t. The intercept b and slope m are statistically estimated from the observations using the maximum likelihood estimation. The proposed POD model building process to evaluate classifier performance is illustrated in Fig. 5.

4. Experimental results

The experimental setup, design, and research methods are presented in this section. The prediction time of drivers lane change classification is used as process parameter. Here the reliability of outcomes of classifiers is examined by using the POD approach. Therefore the POD model building process 5 is implemented. The obtained results and the analysis are discussed in detail.

Fig. 6. SCANeR^{TM} studio, Chair Dynamics and Control, UDuE, Germany.

4.1. Experimental set-up

The driving simulation is performed using $SCANeR^{TM}$ studio driving simulator (Fig. 6). 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.

The driving simulator simulation engine and corresponding input sensors (Fig. 7) aids in decision making. With sensor collected information, driving assistant system (human behavior prediction model) can be established, and finally suggestions/warnings given to driver to control the vehicle's direction, speed, overtake other drivers among others.

The driving environment is a highway-based traffic scenario with four lanes of two directions and simulated traffic environment (Fig. 8). During driving, the participant could perform overtaking maneuver when the preceding vehicle drives slowly. After overtaking the participant is permitted to drive back to the initial lane. The time points of changing lane to left and right are decided by the participant. The lane change/keeping behaviors are illustrated in Fig. 8.

4.2. Data processing

As detailed in Section 2, the algorithms are used to train conventional and modified models. The driving behaviors prediction model based on the classifiers is shown in Fig. 9. It consists of two processes: parameter definition and driving behaviors prediction.

4.3. Result

To demonstrate the POD approach first classical evaluation will be presented. To evaluate the performance of driving behaviors prediction, a common method (Deng & Söffker, 2019) is used, in which the selected evaluation metrics are calculated for the complete driving sequence. Accuracy is one of the commonly used metric due to its simplicity (Hossin & Sulaiman, 2015). A high accuracy can be achieved by correctly classifying the majority class whereas neglecting the minority class. To avoid this known problem, F-measure (FM) representing the harmonic mean between the precision and recall is selected. The measured driving behavior and the estimated driving behavior by the classifier are compared to check the correspondence. The FM of each driving behavior are calculated using Fig. 4 as

$$FM = \frac{2 * precision * recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}. \tag{11}$$

The FM values of the driver using conventional and modified algorithms are shown in Fig. 10.

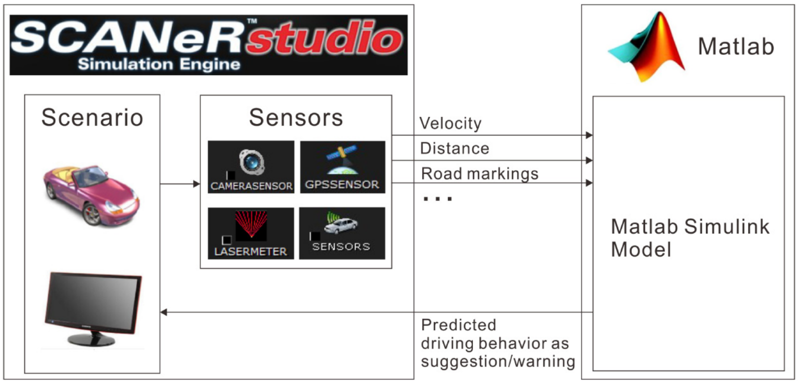


Fig. 7. Sensors for data collection.

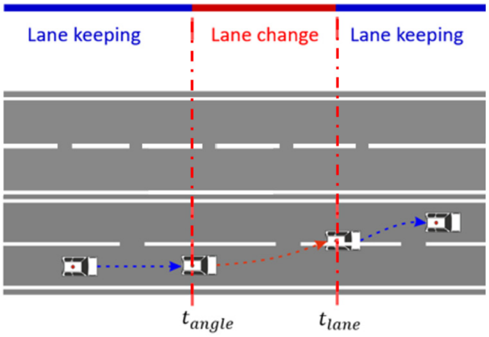


Fig. 8. Lane changing behavior.

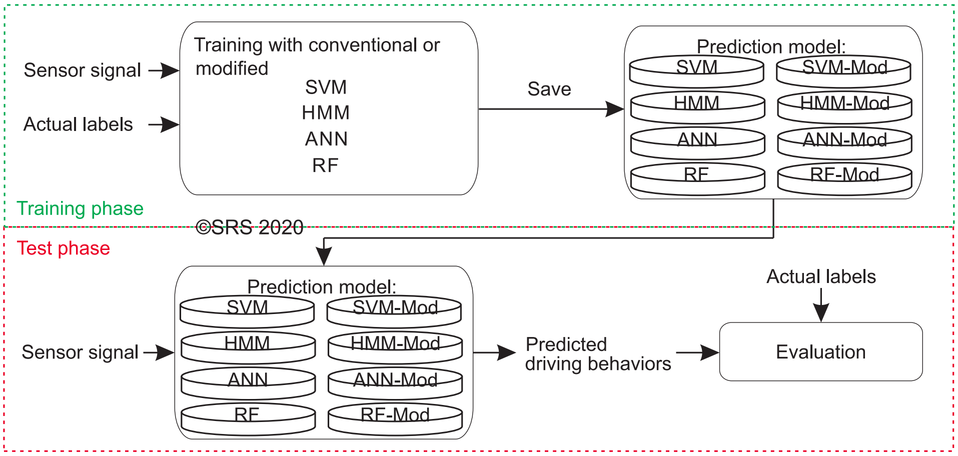


Fig. 9. Test/training model.

It is observed that the modified models for SVM, HMM, and RF are better compared to the conventional model. For ANN, the modified model results are reduced, however, the differences are minimal. Therefore, the overall result considering all situations are still improved using modified algorithms. The best DR vs. FAR values for each classifier is depicted by the ROC graph in Fig. 11.

It should be noted from the results, that both the FM and ROC results are not capable to relate detectability to a process parameter. The POD approach is implemented on the same data in the next section and the aforementioned limitation solved.

4.4. Target response value

The Target-response method as explained in Section 3.1 is used when a relationship between a changing parameter and a changing function or response exist. 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.

To evaluate the predicted lane change performance, each lane change behavior is defined as a separate event. From 7 s 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.

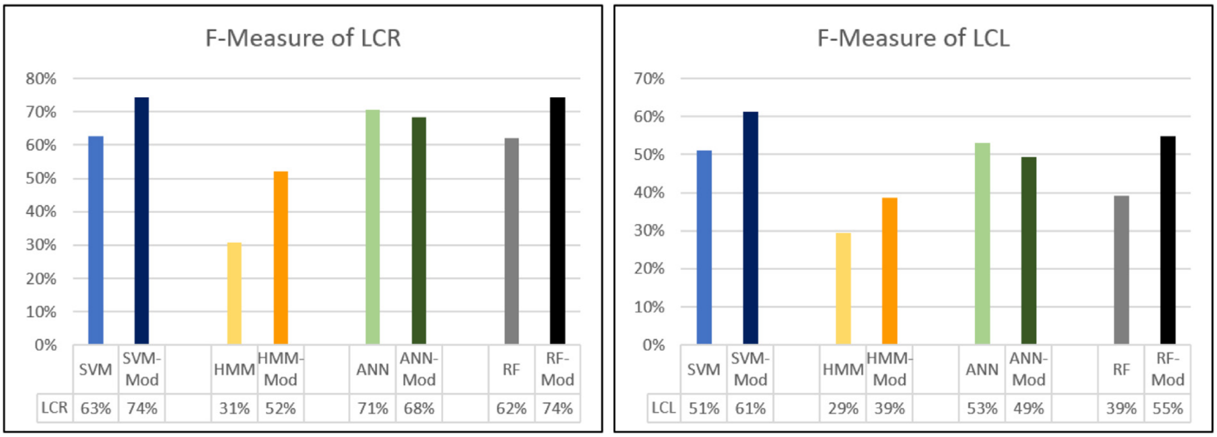


Fig. 10. F-measure results corresponding to each ML algorithm, LCR: lane change to right, LCL: lane change to left.

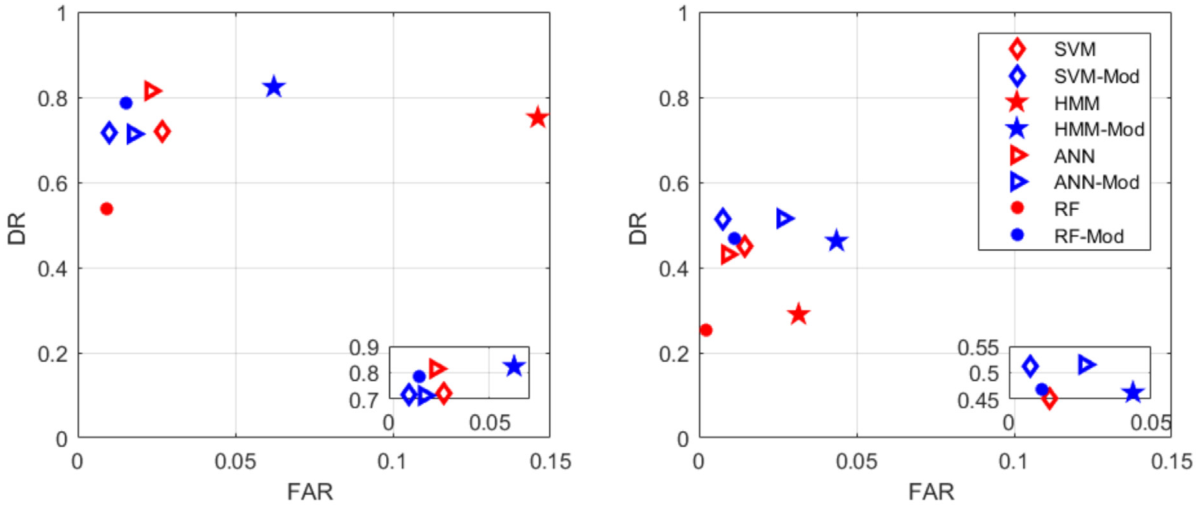


Fig. 11. Best DR/FAR values for classifiers, LCR: lane change to right, LCL: lane change to left.

4.5. POD generation process

Based on the computed response values, the target-response method is utilized in this section to generate POD for the classification 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 results.

Four models comprising combinations of logarithmic and linear scales (Fig. 12) are established to ascertain model described by a straight line and approximately constant variance.

For $X\beta = \sum_i x_i \beta_i$ when X is a row vector and β is a column vector, then

$y = X\beta \rightarrow y_i = \sum_k x_{i,j=k}\beta_{i=k}$ where X is a row vector and y and β are column vectors.

The criteria for a valid model are (Annis, 2020; DOD, 2009)

1 Linearity of the parameters: $E(y_i|X) = x_i\beta$, where x_i is the ith row of X,
2 Uniform variance: $var(y_i|X) = \sigma^2, i = 1, 2, 3, \dots, n$ and
3 Uncorrelated observations: $cov(y_i, y_j|X) = 0, (i \neq j)$.

In this concrete example the model satisfying the above criteria best is the graph with logarithmic abscissa and Cartesian ordinate (model 12 *b*) and hence selected for further analysis. Regression analysis is implemented on model 12 *b* using maximum likelihood estimation as it is better suited for censored data in comparison to known methods like ordinary least squares. The inspection threshold (minimum detectable

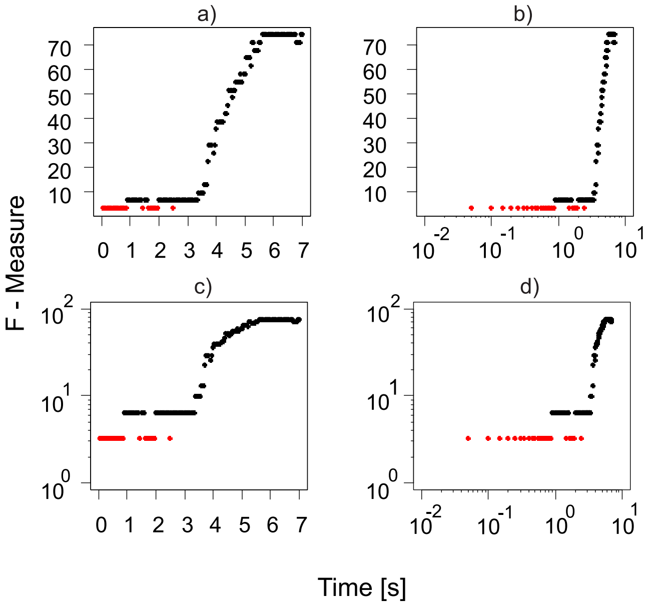


Fig. 12. Regression models a: x vs. y b: log x vs. y c: x vs. log y d: log x. vs. log y.

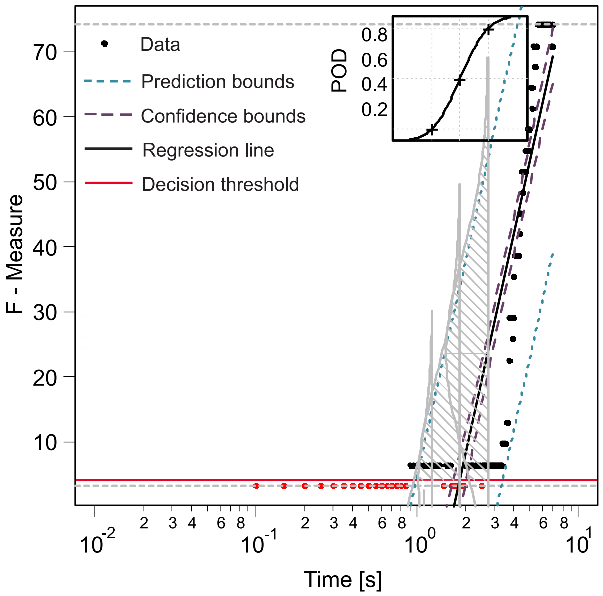


Fig. 13. POD generation process.

data), saturation threshold (maximum inspection threshold), decision threshold (response value below which data is considered as noise), confidence bounds, and prediction bounds are constructed using the formulation from Section 2 as illustrated in Fig. 13.

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. 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.

4.6. Noise analysis

The observed data aggregate the characteristics of the targets signature corrupted by aberrant signals generally referred to as noise. Classical POD methods usually 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 because the corresponding POD is zero. Still using data from Fig. 12, the extracted noise is shown in Fig. 14. 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 deviation σ_{noise} are calculated (Fig. 15). For a Lognormal noise distribution, the false alarm rate is computed as

$$FAR = \int_{y_{th}}^{\infty} \frac{1}{y\hat{\sigma}_{noise}\sqrt{2\pi}} e^{-\frac{(\ln y-\hat{\mu}_{noise})^2}{2\hat{\sigma}^2_{noise}}} dy. \quad (12)$$

The distribution with regards to FAR is illustrated in Fig. 16. That is represented by the shaded red area relative to the selected decision threshold.

From Fig. 16 it becomes evident that for a selected decision threshold (DT), a corresponding unique FAR value exists however the detection probability varies relative to a parameter (here: time). This implies, the premise for the construction of the ROC/PR curve for applications requiring the incorporation of process parameter is deficient. This is because 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 because the size of opposition objects/planes during WWII was irrelevant, but modern applications in the NDT field have demonstrated 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. 17. The FAR values for the selected decision thresholds (DT) are shown in Table 1.

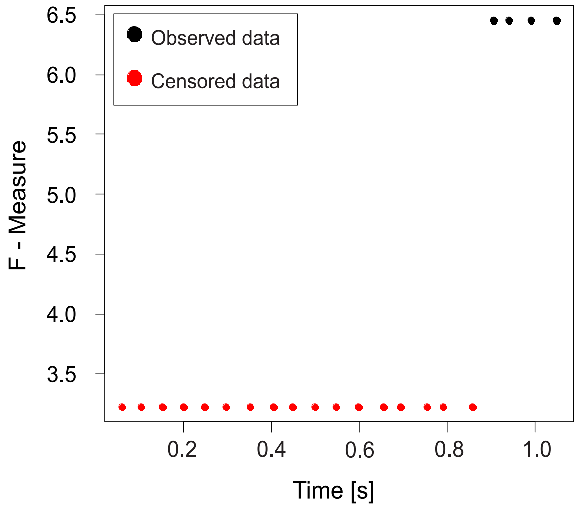


Fig. 14. Noise data.

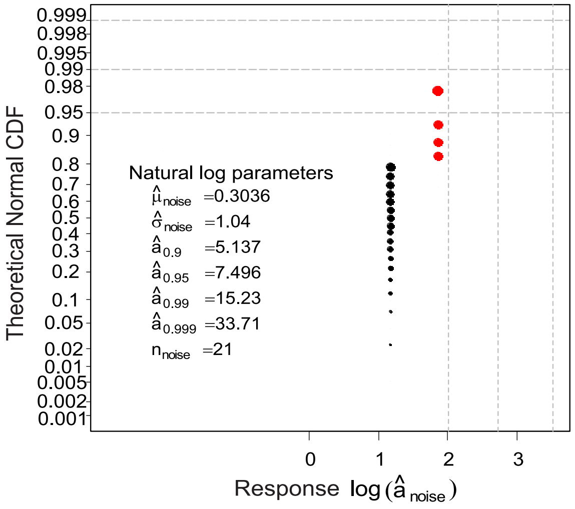


Fig. 15. Noise parameters.

4.7. New measure integrating process parameter

In this section, 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 4.6 are used.

To introduce the novel approach, a single probability point is analyzed. Here the 0.9 probability is used and drawn to intercept POD and confidence curves at the points (+) and (**x**) respectively (shown in Fig. 18). At the point of intersection: the POD, FAR, decision threshold, and the 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. 19.

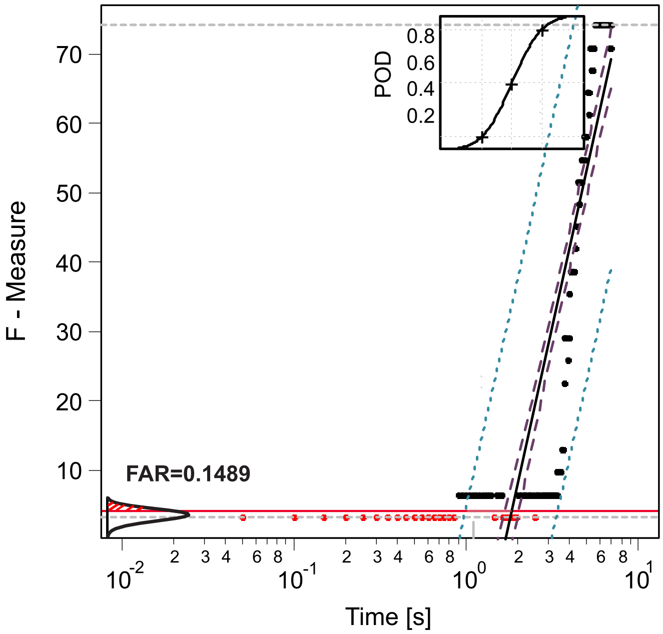


Fig. 16. FAR for a selected decision threshold.

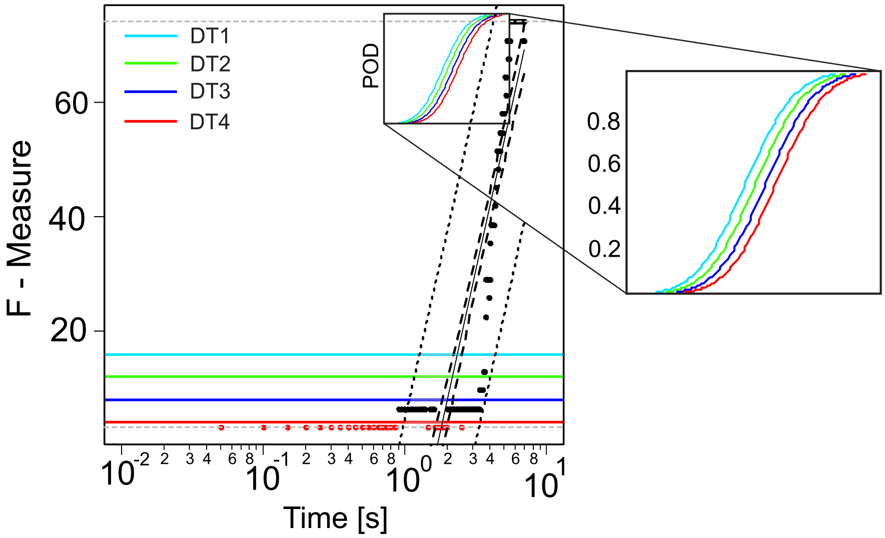


Fig. 17. Relationship between decision threshold (DT) and POD.

Table 1
FAR results for different decision thresholds.

DT [%]	DT1	DT2	DT3	DT4
FAR	0.1489	0.0439	0.0180	0.0088

This insight is helpful because it provides an additional measure to incorporate and assess the effect of process parameters on the evaluation process. In Fig. 19 the evaluation is undertaken for the 0.9 probability. If the complete range from 0 to 1.0 is examined, the relation of every probability point can be evaluated. The resulting relation corresponding to the entire probability range utilizing the developed method is indicated in Fig. 20.

The entire POD range contains detection probabilities from 0.0 DR up to 1.0 DR. From the illustrated relations two points can be analyzed. 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.

5. Central outcome: comparison of classifiers using POD

Based on the developed POD approach the ability of eight different classifiers to predict driver lane change behavior is examined. The predicted driving behavior as a function of target response at each recognition point for the same selected threshold is shown in Figs. 21 and 22. In Figs. 21 and 22, the FM values for left and right lane at each recognition point is shown respectively. Based on the FM data the corresponding POD curves are generated for left lane change (Fig. 23) and right lane change (Fig. 24).

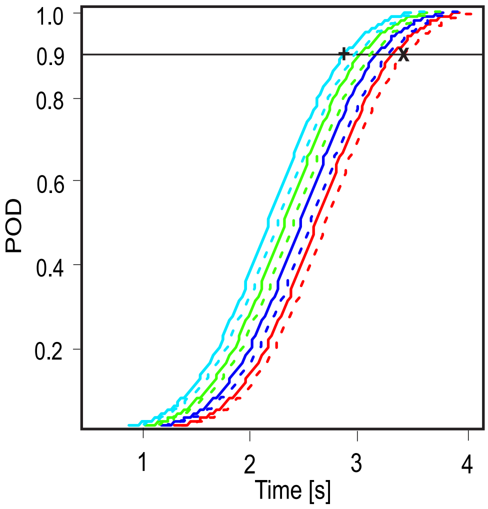


Fig. 18. Detection at 90% and 90/95 level, +: 90% POD, **x**: 90/95 POD.

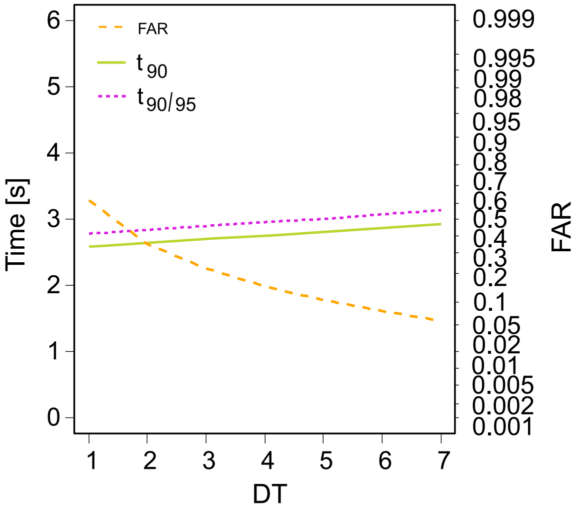


Fig. 19. Evaluation measure incorporating process parameter, t_{90}: 90% POD, $t_{90/95}$: 90/95 POD.

The illustration allows to evaluate prediction capabilities of classifiers incorporating a process parameter (here: the time [in sec.] before the impending event). 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. This is because the algorithm is able to predict the complete lane change behavior faster therefore it has better prediction capabilities.

The Prediction Time Ratio (PTR) which is a ratio of the classifier prediction time to the complete lane change can be calculated as

$$PTR = \left(\frac{Total\ lane\ change\ time - Classifier\ prediction\ time}{Total\ lane\ change\ time}\right) \times 100\ \%. \tag{13}$$

The higher PTR values represent better results compared to lower PTR values. The 90/95 POD values and the comparison between the conventional and modified classifiers for left and right lane changes are illustrated in Table 2. The corresponding PTR values are shown in Table 3.

For this specific example, the experimentally obtained and analyzed results can be summarized as follows,

In estimating left lane change:

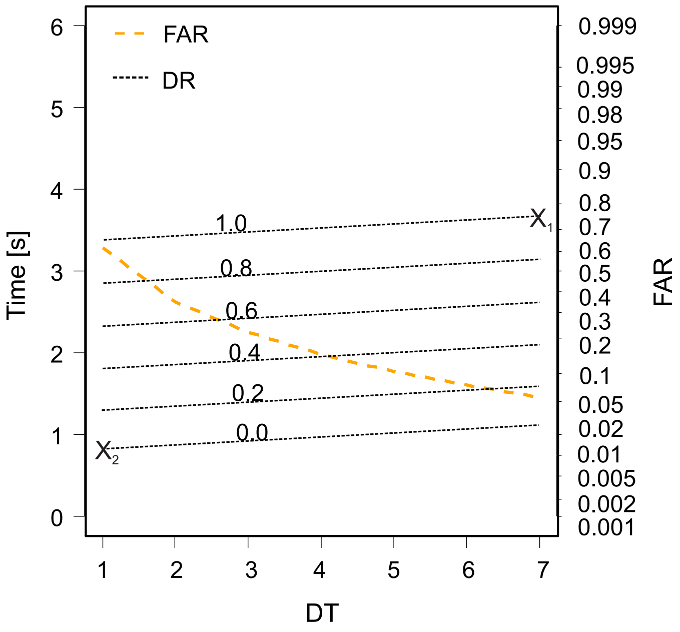


Fig. 20. Parametric evaluation measure.

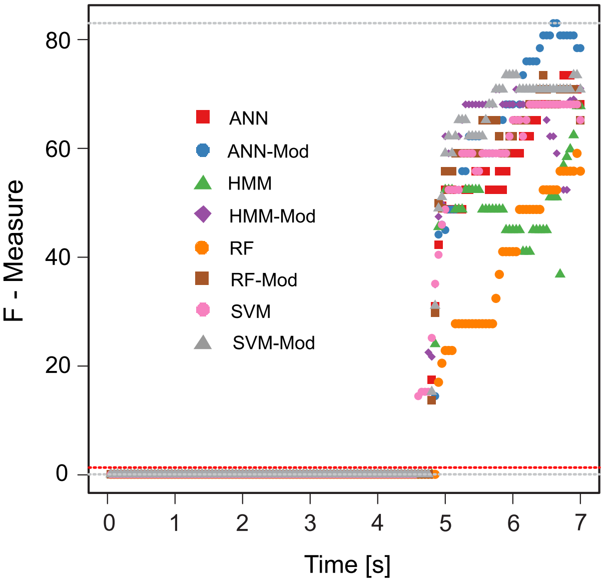


Fig. 21. Left lane change classifier data.

1 Conventional SVM produces best results (4.781 s) and conventional HMM produces worst results (5.776 s). This implies conventional SVM is capable to predict the lane change 2.219 s (7−4.781) before the actual maneuver representing a PTR of 31.7% ($\frac{7-4.781}{7} \times$ 100%) earlier compared to 17.486% ($\frac{7-5.776}{7} \times$ 100%) before complete maneuver time for conventional HMM.

2 The use of prefilters (ANN/HMM/RF/SVM-Mod) to tune parameters with the aim to optimize the FM performance of conventional algorithms (Deng & Söffker, 2019) 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:

1 Modified ANN produces best results (4.565 s) and conventional HMM producing worst results (5.138 s). This is corroborated by

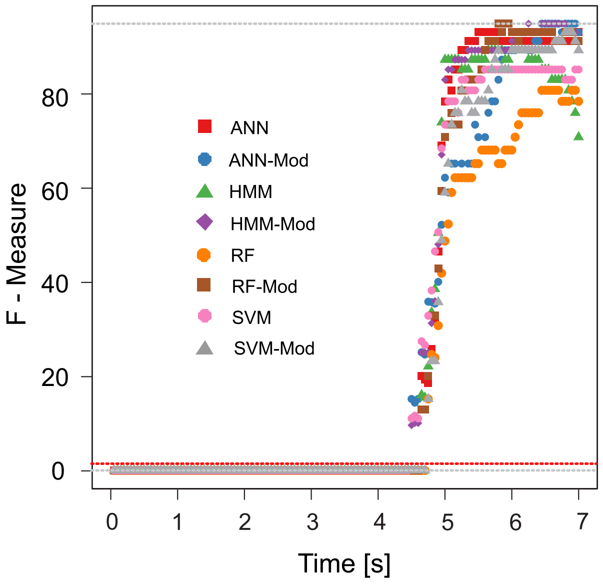


Fig. 22. Right lane change classifier data.

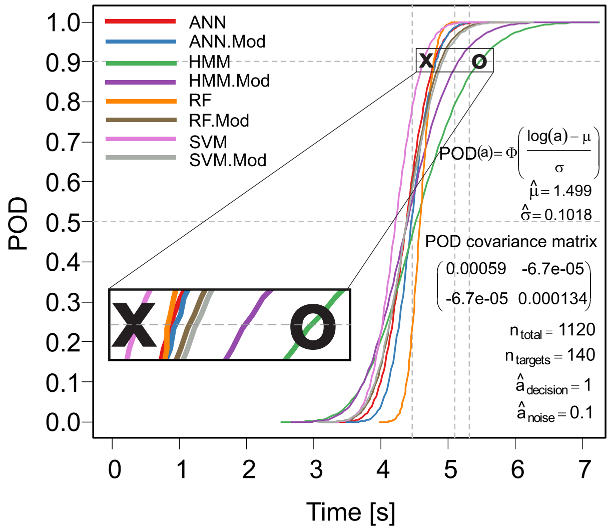


Fig. 23. Left lane change POD results, **x**: best **o**: worst.

Table 2
POD results and comparison between conventional and modified classifiers.

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^{w}	4.565^{b}
HMM-Mod	5.464^{b}	4.871^{b}
RF-Mod	5.076^{w}	4.952^{w}
SVM-Mod	5.113^{w}	4.979^{w}

Legend- Mod: modified
b: improved results
w: worse results

the PTR values of 34.786% and 26.6% for modified ANN and conventional HMM respectively.

2 The use of prefilters (ANN/HMM/RF/SVM-Mod) to tune parameters with the aim to optimize the performance of conventional algorithms results in two improved POD (ANN-Mod and HMM-Mod) and two worsened POD (RF-Mod and SVM-Mod) values.

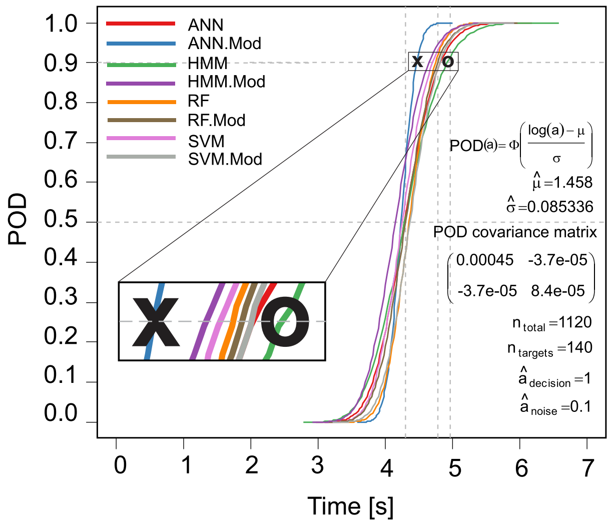


Fig. 24. Right lane change POD results, **x**: best **o**: worst.

Table 3
Prediction time as a percentage of total lane change time.

Classifier	Left lane PTR [%]	Right lane PTR [%]
ANN	29.157	28.557
HMM	17.486**	26.6**
RF	30.343	30.057
SVM	31.7*	31.1
ANN-Mod	29.014	34.786*
HMM-Mod	21.943	30.414
RF-Mod	27.486	29.257
SVM-Mod	26.957	28.871

Legend- Mod: modified
*: best results
**: worst results

Accordingly for this example it can be concluded that the best classifier is conventional SVM for left lane change prediction whereas modified ANN is best for right lane change prediction. Conventional HMM produces worst results in both instances. An interesting observation from the results is that, the use of prefilters to optimize the performance of conventional algorithms is partially achieved for right lane change prediction while worst results are obtained for left lane change prediction.

5.1. Improvement of classifier detection capability

Comparison of the introduced method with improved approach is made in this section. Applying the approach introduced it can be generalized that the generated POD is not unique but dependent on predictor response or response value. To improve the POD results, a strategy involving the utilization of a crisp target response evaluation value will be used. The F measure is calculated using the precision and recall. The F measure depends on test and population. The detection rate on the other hand depends solely on the test dataset. The DR response value can thus be selected and used for evaluation to produce an improved POD though the FM provides a better generalization. The DR is used as a response value. The 90/95 POD values for all eight classifiers using the introduced approach are shown in Table 4. The corresponding PTR values are shown in Table 5.

The results show an improvement in the POD values for all classifiers in comparison to when FM is used as response value. This can

Table 4
POD results using DR values and comparison between conventional and modified classifiers.

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^{b}	3.679^{w}
HMM-Mod	0.6355^{b}	0.5748^{b}
RF-Mod	2.883^{b}	3.181^{b}
SVM-Mod	3.300^{w}	3.485^{w}

Legend- Mod: modified
b: improved results
w: worse results

Table 5
Prediction time as a percentage of total lane change time.

Classifier	Left lane PTR [%]	Right lane PTR [%]
ANN	56.186	52.5
HMM	79.443	82.8
RF	39.729**	52.086
SVM	65.243	57.843
ANN-Mod	83.671	47.443**
HMM-Mod	90.921*	91.789*
RF-Mod	58.814	54.557
SVM-Mod	52.857	50.214

Legend- Mod: modified
*: best results
**: worst results

be observed from the PTR values. For left lane change, using DR as response value produces highest PTR value of 90.921% and lowest PTR value of 39.729% (see Table 5) compared to highest PTR value of 30.343% and lowest PTR value of 17.486% when FM is used as response value (see Table 3). For right lane change, using DR as response value produces highest PTR value of 91.789% and lowest PTR value of 47.443% (see Table 5) compared to highest PTR value of 34.786% and lowest PTR value of 26.6% when FM is used as response value. This effective strategy can be implemented to generate improved POD results. It is also observed that the usage of prefilters to tune parameters with the aim to optimize the performance of conventional algorithms generally works to improve the classifiers POD-related performance. 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.

From the results shown in Tables 2, 3, 4, and 5 it can be clearly seen that the introduced POD approach helps to evaluate detailed problem-related questions regarding physics-based process parameters (here: prediction evaluation of classifiers with respect to timely distance to the event to be predicted). The presented approach provides a new measure to evaluate the effects of process parameters on classification results to the engineering-oriented machine learning community. The introduced approach can be generalized for other classification models like Long Short-Term Memory (LSTM) networks (a type of recurrent neural network). Given that LSTM is designed for sequence problems and capable of handling temporal dynamic behavior (Sherstinsky, 2020), the POD approach come in handy as it is process parameter dependent and therefore the reliability at each stage of the sequence can be thoroughly evaluated.

6. Conclusion

In this contribution a new assessment on the performance of Machine Learning-based classification approaches using POD approach is presented. This is needed because the evaluation process of known approaches like ROC/PR are non-parametric and hence not suitable to evaluate the effect of process parameters on the classification results. It can be shown that a selected decision threshold correspond to a unique FAR however several DR values exist. This implies that the ROC and PR measures cannot be used for applications which requires detectability to be related to a process parameter. This missing link leads to the non acceptance of machine learning-related algorithms in safety critical context. This deficiency is addressed using the newly established POD measure. The new approach is developed and additionally demonstrated on experimental data using human driving behaviors. The target-response method is implemented as a new analysis and certification tool for classifiers permitting the comparison of different ML algorithms. Noise analysis procedure is additionally introduced permitting trade-off between DR, FAR, and process parameter. The introduced approach provides a comprehensive interpretation of the quality of arbitrary classification models by incorporating a process parameter in the evaluation process. This allows a new evaluation and certification standard for machine learning approaches.

CRediT authorship contribution statement

Daniel Adofo Ameyaw: Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing – original draft, Writing – review & editing. **Qi Deng:** Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing – original draft, Writing – review & editing. **Dirk Söffker:** Conception and design of study, Analysis and/or interpretation of data, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

All authors approved the version of the manuscript to be published.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.mlwa.2021.100220.

References

Ali, M. S., Miah, M. S., Haque, J., Rahman, M. M., & Islam, M. K. (2021). An enhanced technique of skin cancer classification using deep convolutional neural network with transfer learning models. *Machine Learning with Applications*, *5*, Article 100036.

Ameyaw, D. A., Deng, Q., & Söffker, D. (2019). Probability of detection (POD)-based metric for evaluation of classifiers used in driving behavior prediction. In *Annual conference of the PHM society, vol. 11 no. 1*.

Ameyaw, D. A., Rothe, S., & Söffker, D. (2020). A novel feature-based probability of detection assessment and fusion approach for reliability evaluation of vibration-based diagnosis systems. *Structural Health Monitoring*, *19*(3), 649–660.

Annis, C. (2020). Statistical best-practices for building probability of detection (POD) models. In *R Package Mh1823, Vol. 2*. (4.4), Statistical Engineering.

Bowyer, K., Kranenburg, C., & Dougherty, S. (2001). Edge detector evaluation using empirical ROC curves. *Computer Vision and Image Understanding*, *84*(1), 77–103.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2019). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences*, *557*, 317–331.

Cook, J., & Ramadas, V. (2020). When to consult precision-recall curves. *The Stata Journal*, *20*(1), 131–148.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Deng, Q., & Söffker, D. (2018). Improved driving behaviors prediction based on Fuzzy Logic-hidden Markov model (FL-HMM). In *2018 IEEE intelligent vehicles symposium* (pp. 2003–2008). IEEE.

Deng, Q., & Söffker, D. (2019). Classifying human behaviors: Improving training of conventional algorithms. In *2019 IEEE intelligent transportation systems conference* (pp. 1060–1065). IEEE.

Deng, Q., Wang, J., & Söffker, D. (2018). Prediction of human driver behaviors based on an improved HMM approach. In *2018 IEEE intelligent vehicles symposium* (pp. 2066–2071). IEEE.

DOD, U. (2009). Department of defense handbook: nondestructive evaluation system reliability assessment. In *MIL-HDBK-1823A* (2nd ed.). Washington, DC: DOD.

Gandossi, L., & Annis, C. (2010). Probability of detection curves: Statistical best-practices. In *ENIQ Report, office for official publications of the European communities, vol. 41*.

Gao, W., Wang, L., Jin, R., Zhu, S., & Zhou, Z.-H. (2016). One-pass AUC optimization. *Artificial Intelligence*, *236*, 1–29.

Georgiou, G. (2007). PoD curves, their derivation, applications and limitations. *Insight-Non-Destructive Testing and Condition Monitoring*, *49*(7), 409–414.

Ginzel, E. (2006). Introduction to the statistics of NDT. *E-Journal of Nondestructive Testing*, *11*(5), 4–7.

González-Carrasco, I., Jiménez-Márquez, J. L., López-Cuadrado, J. L., & Ruiz-Mezcua, B. (2019). Automatic detection of relationships between banking operations using machine learning. *Information Sciences*, *485*, 319–346.

Harding, C. A., Hugo, G. R., & Bowles, S. J. (2009). Application of model-assisted POD using a transfer function approach. In *AIP conference proceedings, vol. 1096* (1), (pp. 1792–1799). American Institute of Physics.

Hossin, M., & Sulaiman, M. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 1.

Khan, J. Y., Khondaker, M. T. I., Afroz, S., Uddin, G., & Iqbal, A. (2021). A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications*, *4*, Article 100032.

Knopp, J. S., Aldrin, J. C., Lindgren, E. A., & Annis, C. (2007). Investigation of a model-assisted approach to probability of detection evaluation. In *AIP conference proceedings, vol. 894* (1), (pp. 1775–1782). American Institute of Physics.

Kukkala, V. K., Tunnell, J., Pasricha, S., & Bradley, T. (2018). Advanced driver-assistance systems: A path toward autonomous vehicles. *IEEE Consumer Electronics Magazine*, *7*(5), 18–25.

Kutner, M. H., Nachtsheim, C. J., Neter, J., Li, W., et al. (2005). *Applied linear statistical models, vol. 5*. Irwin, New York: McGraw-Hill.

Lin, G.-M., Nagamine, M., Yang, S.-N., Tai, Y.-M., Lin, C., & Sato, H. (2020). Machine learning based suicide ideation prediction for military personnel. *IEEE Journal of Biomedical and Health Informatics*, *24*(7), 1907–1916.

Lorente, D., Aleixos, N., Gómez-Sanchis, J., Cubero, S., & Blasco, J. (2013). Selection of optimal wavelength features for decay detection in citrus fruit using the ROC curve and neural networks. *Food and Bioprocess Technology*, *6*(2), 530–541.

Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S., & Coello, C. A. C. (2013). A survey of multiobjective evolutionary algorithms for data mining: Part I. *IEEE Transactions on Evolutionary Computation*, *18*(1), 4–19.

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How many trees in a random forest? In *International workshop on machine learning and data mining in pattern recognition* (pp. 154–168). Springer.

Ozenne, B., Subtil, F., & Maucort-Boulch, D. (2015). The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of Clinical Epidemiology*, *68*(8), 855–859.

Peng, R. D., & Hengartner, N. W. (2002). Quantitative analysis of literary styles. *The American Statistician*, *56*(3), 175–185.

Qiu, S., Rachedi, N., Sallak, M., & Vanderhaegen, F. (2017). A quantitative model for the risk evaluation of driver-ADAS systems under uncertainty. *Reliability Engineering & System Safety*, *167*, 184–191.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Rong, H., Teixeira, A., & Soares, C. G. (2020). Data mining approach to shipping route characterization and anomaly detection based on AIS data. *Ocean Engineering*, *198*, Article 106936.

Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*, Article 132306.

Theissler, A., Pérez-Velázquez, J., Kettelgerdes, M., & Elger, G. (2021). Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry. *Reliability Engineering & System Safety*, *215*, Article 107864.

Thompson, R. B., Brasche, L. H., Forsyth, D., Lindgren, E., Swindell, P., & Winfree, W. (2009). Recent advances in model-assisted probability of detection. In *4th European-American Workshop on Reliability of NDE*.Berlin.

Volker, A., Dijkstra, F., Terpstra, S., Heerings, H., & Lont, M. (2004). Modeling of NDE reliability: development of a POD generator. In *Proceedings of the 16th World Conference on Nondestructive Testing*. Montreal, Canada, August.