Risk-based Evaluation of ML Classification Methods Used for Medical Devices

Martin Haimerl (✉ Martin.Haimerl@hs-furtwangen.de)
Furtwangen University

Christoph Reich
Furtwangen University

Research Article

Keywords: Classification, Risk Management, Risk-based Metrics, Decision Theory, Medical Devices

Posted Date: September 13th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3317894/v1

Additional Declarations: No competing interests reported.

Risk-based Evaluation of ML Classification Methods Used for Medical Devices

Martin Haimerl[1]

Christoph Reich[1]

[1]Furtwangen University of Applied Sciences, Furtwangen, Germany

Corresponding Author: Martin Haimerl,

Martin.Haimerl@hs-furtwangen.de

Abstract

Background: In the future, more and more medical devices will be based on machine learning (ML) methods. For such medical devices, the rating of risks is a crucial aspect and should be considered when evaluating their performance. This means that an integration of risks and their associated costs into the corresponding metrics should be taken into account. This paper addresses three key issues towards a risk-based evaluation of ML-based classification models.

Methods: First, it analyzes a selected set of scientific publications for determining how often risk-based metrics are currently utilized in the context of ML-based classification models. Second, it introduces an approach for evaluating such models where expected risks and associated costs are integrated into the corresponding performance metrics. Additionally, it analyzes the impact of different risk ratios on the resulting overall performance. For this purpose, an artificial model was used which allows to easily adapt key parameters. Third, the paper elaborates how such risk-based approaches relate to regulatory requirements in the field of medical devices. A set of use case scenarios were utilized to demonstrate necessities and practical implications, in this regard.

Results: With respect to the first research question, it was shown that currently most scientific publications do not include risk-based approaches for measuring performance. For the second topic, it was demonstrated that risk-based considerations have a substantial impact on the outcome. The relative increase of the resulting overall risks can go up to 198%, i.e. the risk value almost triples, when the ratio between different types of risks (risk of false negatives in comparison to false positives) goes down/up to 0.1 or 10.0. As discussed within the third research question, this situation typically represents a case where the risk increases one level in the corresponding risk matrix. Based on this, it was demonstrated that differences in parameter settings lead to a substantially different behavior when risk factors are not addressed properly.

Conclusion: In summary, the paper demonstrates the necessity of a risk-based approach for the evaluation of ML-based medical devices, develops basic steps towards such an approach, and elaborates consequences which occur, when these steps are neglected.

Keywords: Classification; Risk Management; Risk-based Metrics; Decision Theory; Medical Devices.

1 Background

Machine learning (ML) is a revolutionary technology which is more and more applied in concrete medical applications (cf. (1–3)). In specific tasks like diagnosis of diseases, e.g. skin cancer or retinal diseases, ML techniques achieve an equivalent or even better accuracy in comparison to human experts (2, 4). Such results indicate that the utilization of ML-based methods in actual clinical applications is promising and there already is a series of ML-based medical devices which were successfully placed on the market (5). However, the clinical impact of the used devices has to be clearly demonstrated for the particular use case. Thus, a thorough evaluation with respect to the performance of the ML algorithms and their effect in the actual clinical environment has to be performed. For example, the requirements from the medical device regulation (MDR) (6) have to be fulfilled, before the device can be placed on the European Union (EU) market. In the future, also the proposed AI Act (7) has to be applied. The conformity with these regulations is usually proven by means of the harmonized standards associated with them. For performing risk management in the context of medical devices, the ISO 14971 (8) is the appropriate standard. Additionally, the technical report ISO/TR 24971 (9) provides more detailed guidance for the application of (8). But, neither the MDR (6) nor (8, 9) contain specific information for AI/ML-based devices. Thus, a dedicated framework for addressing risk management in these cases is still missing.

The basic aim of the regulations is that the devices achieve a level of safety and performance which is appropriate for the clinical application. This includes a thorough analysis of potential risks and their associated impact as well as the clinical performance of the device with respect to the specific application and its context. In general, risk refers to an uncertain outcome. In particular, risks are related to potential harms and are defined as a combination of a certain likelihood, i.e. probability of occurrence, and a severity, i.e. magnitude of harm This is also represents the definition in ISO 14971 (8). The intent behind risk management is to identify, evaluate, analyze, assess, and mitigate potential product issues. According to (6), risks have to be reduced as far as possible unless avoidance of further risk improvements does not have an adversarial effect on the risk-benefit relationship. Finally, the risks have to outweigh the benefits. Thus, it is crucial to evaluate the clinical outcome of a device as the central criterion. For ML-based devices, this means that performance measures should be established which include such factors. The associated risks are one major component in this regard. Additionally, the achieved benefits are important factors. To a certain degree, they can be considered as negative risks. Pure error or accuracy rates are not sufficient for evaluating the clinical performance of the device.

Currently, it seems that most of the scientific publication use standardized performance metrics, which basically focus on accuracy-based assessments to validate and test their ML models. This means that only the differences between the predicted results and the values from the reference data set (training, validation or test data sets) are compared, in particular, when considering supervised ML methods. For classification tasks, this includes metrics like accuracy, precision, sensitivity/recall, $F1$ score, Matthews Correlation Coefficient (MCC), or Area under the ROC Curve ($AUROC$) (see e.g. (10) for an overview about applicable metrics). For example, this can be recognized in the preprint (11), where more than 70 medical image experts systematically analyzed requirements regarding the evaluation of machine learning models, e.g. for image-level classification tasks. Only very limited references were included, where risks, costs, or benefits were included in the metrics, e.g. in terms of net benefit (12) or expected costs (13). Additionally, the weighted kappa statistic and the F_β score were mentioned as options to integrate weightings. But, concrete advises how to determine and integrate appropriate weights were not given in (11). Instead, most of the recommendations were based on the application of standard metrics, like the ones mentioned above. The hypothesis that most recent scientific publications do not systematically address risk factors within the evaluation of ML models was one major goal of the analysis performed within this paper.

In the mentioned standardized metrics, only the number of errors is taken, when considering classification tasks, but not the impact of the different type of errors. For example, a false negative ("missed diagnosis") can have a substantially different clinical effect than a false positive ("false alarm"), when considering diagnostic applications. For example, a false positive within a cancer screening may have some harm (e.g. feeling of insecurity, additional tests with potential harm). But, the harm in these cases is often considerably lower than the harm of false positives. A missed diagnosis may leed to substantial progression of the disease and eventually also to a lethal outcome. These are important issues since the associated risk impact usually goes in contrary directions and thus need to be balanced out in a dedicated way.

The standard performance metrics, which are used in many publications, do not include a dedicated assessment with regards to the risks and their clinical impact of a particular use case. Only the deviation / consistency rate between the training samples and the prediction of the models is optimized. Implicitly, the performance metrics assume some kind of neutral situation, where a certain balancing of the relationship between false positives and false negatives is given. They basically reflect the relationships as they are represented in the used data sets, but not the associated relationship of risks. Usually, the balancing of data sets, e.g. providing the same number of false positives and false negatives, is a recommendation to achieve a certain level of adjustment since one type of error often is predominant (11). However, this does only represent a standardized rule lacking a dedicated adaption to a particular use case. Of course, there are further important aspects which have to be considered in the quality assessment of ML-based techniques, like data quality or uncertainty factors, e.g. in terms of confidence intervals for the results (14).

For utilization of ML-based techniques in medical devices, such risk factors have to be included to consequently follow the rules given by the regulations and standards, like (6) and (8). Otherwise, the reduction of the risks and the optimization of clinical benefits remains deficient. One approach to achieve this for ML based classification tasks is an appropriate adjustment of threshold parameters, after the training procedure. However, the risk factors are not fully integrated into the development and evaluation of the models, in this case. To achieve this, in a comprehensive way, the different impact of false positives and false negatives has to be integrated into the performance metrics, when evaluating the results of binary classification problems. For example, in (15) it was demonstrated, that a cost-effectiveness analysis can lead to very different results, when considering actual costs for different treatment options. This was shown for a concrete medical application, i.e. proximal caries detection, where the analysis focused on a comparison between an ML-based and a conventional approach (15).

Thus, the selection of the best model should be performed in terms of the best decision not only with respect to measures of deviation. It should be addressed in terms of the best clinical outcome, the strongest reduction of costs, and the risks for the specific application. Since the likelihood of risks and its corresponding harm is usually not given exactly, this can only be achieved in a probabilistic manner, i.e. as an optimization of the expected costs and benefits when applying the model. Such approaches are linked to the field of decision theory (16). An application specific utility function has to be defined and optimized to achieve the best outcome. This approach can be combined with a risk analysis and its associated risk factors, e.g. as described in (17, 18). In particular, this had been applied to classification problems in medical applications (19, 20, 12) as well as to medical decision making in a general context (21). Additionally, it was proposed as a basic rationale for optimizing ML models (22). This approach converts the construction of the ML model into a process for finding an optimal decision rule based on probabilities and weights (i.e. costs or utilities) of the corresponding risks and benefits.

The current paper follows this approach for evaluating the performance of ML models based on risk profiles of the specific clinical application and integrating such methods into the development of ML-based medical devices. It analyses the impact, that results from variations in risk profiles. The paper focuses on binary classification tasks and subsequently on the evaluation of the outcome in terms of appropriate performance metrics. Other important quality factors, like data quality, uncertainty assessment, or also the interpretability of the models (see e.g. (7) for relevant aspects), are not addressed within this paper, in a dedicated way. Instead, the paper aims at clarifying the relationship between risk management requirements and performance assessment. For this purpose, it includes the analysis of the following three core topics:

- First, the hypothesis was analyzed that current scientific papers about using ML in medical applications often only use standardized performance metrics without including the (clinical) impact of application-specific risks. This was addressed by a research of recent literature about ML-based classification techniques and their use in medical applications. This was not addressed using a comprehensive survey. Instead, an exemplary literature research was utilized, which analyzes the outcomes according to a sample of articles obtained for a given time frame . See sections 2.1 for the definition of the study and 3.1 for the results.
- Second, a performance assessment was described and applied which is based on assigning dedicated costs / weights to the particular types of errors in a binary classification task. This was demonstrated using an artificial model representing the particular amount of errors. A model was developed which achieves a risk-based evaluation of ML-based classification models. The main goal of this analysis was to determine the impact of different risk ratios on the resulting performance of the model – see sections 2.2 and 3.2.

- Third, the integration of the overall results were assessed in relation to the requirements given by the corresponding standards and regulations, in particular the MDR (7), the proposed AI Act (6), ISO 14971 (8) as the standard for risk management in medical devices, and the technical report ISO/TR 24971 (9) which provides more concrete guidance for implementing the risk management process. For this purpose, a set of use scenarios was utilized to demonstrate the impact of the particular settings on the evaluation of the ML-based models – see sections 2.3 and 3.3.

Preliminary results for the second of these topics were presented in (23). This included a basic model for assessing the impact of risk factors on the outcome of ML-based classification methods. The analysis was substantially extended in this new paper with respect to each of the research questions described above.

2 Methods

The following sections describe the basic methodology as it was applied in this paper for each of the three topics. The results are presented later in the corresponding sections of chapter 3.

2.1 Topic A – Utilization of risk-based performance metrics in recent publications

As a first step, the hypothesis was addressed that most scientific publications about machine learning techniques only apply standardized metrics and do not include use-case specific costs, benefits, or risk factors into their assessment of model performances. This analysis was restricted to concrete use cases and studies in the field of medical applications, where binary classification was a main focus of the publication. For this purpose, a literature research was performed in pubmed (https://pubmed.ncbi.nlm.nih.gov/) including the most recent publication in this field. The goal was to determine the percentage of articles which include such considerations by using this exemplary search. It aimed to analyze how many of the publications contained risk-based considerations for the evaluation of the models, in this particular sample of articles. The following search term was used: *"machine learning" classification (performance OR evalua* OR assess*)* *metric**, where the search terms could appear in any fields. The first two parts were included to select ML-based classification tasks. The remaining part narrowed the search to cases where an assessment based on performance metrics was performed. In pubmed, the different parts of the search terms were combined by an AND-operator, i.e. each particular search term needs to be met. Filters for *free full text* and *in the last 1 year*, i.e. previous year starting from the date of the search, were added to restrict the search to the most recent and freely accessible publications. This was not considered as a major restriction since it still represents a valid cross-sectional sample of articles. Finally, only papers in *English* were selected using another pubmed filter option.

The identified articles were analyzed starting from the most recent towards the more antecedent publications until a number of 30 papers was included into the analysis. The following exclusion criteria were used to only focus on relevant publications.

Exclusion criteria for literature research:

- The main focus / task of the paper was not a direct medical application and/or did not focus on a dedicated clinical study / use case. Based on this, publications from other domains, surveys / systematic reviews, abstract presentation of methods without use case, etc. were excluded.

- Binary classification was not the focus of one of the main endpoints in the study. For borderline cases, where a binary classification results were reported within a multiclass classification task, we restricted our search results to cases, where only a limited number of classes (up to 5) were addressed and the performance of the single classes was a main outcome.

  Remark: The rationale behind this selection was that for multiclass problems with many classes the assessment of risks is even more remote. We wanted to focus on applications where the inclusion of risk factors would be more obvious.

- The used performance metrics were listed in the paper and described in a way, that they can be judged appropriately.

Based on these criteria, the literature search provided a random sample / cross section of recent publications in this field which was further analyzed regarding the used performance metrics for the binary classification task. In particular, this included the following metrics, which are based on the numbers of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) in the results of the binary classification task. Basically, the metrics listed in Tab. 1 were documented within our study.

Tab. 1. Standard performance metrics typically used for ML-based classification tasks. It is assumed that the of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) are given. See (10) for more details about the definition and utilization of these metrics.

General / overarching definitions	
Number of actual positive cases: $\boldsymbol{P = TP + FN}$	**Number of actual negative cases:** $N = TN + FP$
Number of predicted positive cases: $\boldsymbol{PP = TP + FP}$	**Number of predicted negative cases:** $PN = TN + FN$
Total Population: $\boldsymbol{Pop = P + N}$	**Prevalence:** $Prev = \frac{P}{P + N} = \frac{P}{Pop}$
Metrics documented in the literature research within this study	
Sensitivity / Recall / True Positive Rate: $\boldsymbol{TPR = \frac{TP}{P}}$	**Specificity / True Negative Rate:** $TPN = \frac{TN}{N}$
Accuracy: $\boldsymbol{Acc = \frac{TP + TN}{TP + FP + TN + FN}}$ **or equivalently Error rate:** $\boldsymbol{Err = 1 - Acc}$	**Balanced Accuracy,** i.e. accuracy after balancing of positive / negative test samples / class members: $BA = \frac{TPR + TNR}{2}$
Precision / Positive Predicted Value: $\boldsymbol{PPV = \frac{TP}{PP}}$	**Negative Predictive Value:** $NPV = \frac{TN}{PN}$

F_1-Score: $F1 = 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR}$	**other F_β-Scores:** $F\beta = (1 + \beta^2) \cdot \frac{PPV \cdot TPR}{\beta^2 \cdot PPV + TPR}$
Matthews Correlation Coefficient: $MCC = \sqrt{TPR \cdot TNR \cdot PPV \cdot NPV} -$ $\sqrt{(1 - TPR) \cdot (1 - TNR) \cdot (1 - PPV) \cdot (1 - NPV)}$	**Geometric Mean:** $MCC = \sqrt{TPR \cdot TNR}$
Measures which include not single models (fixed threshold) **but multiple variations of thresholds**	
Receiver Operating Characteristics (ROC) Curve, i.e. plot of FPR (on x axis) vs. TPR (on y axis).	**Precision-Recall Curve (PRC),** i.e. plot of recall / TPR (on x axis) vs. precision / PPV (on y axis).
Area under the ROC Curve: $AUROC = \int_0^1 ROC(x)\, dx$ as the integral over the function $ROC(x)$ described by the ROC Curve	**Area under the PRC Curve:** $AUPRC = \int_0^1 PRC(x)\, dx$ as the integral over the function $PRC(x)$ described by the PRC Curve
Measures for comparison of two predictions	
(Cohen's) Kappa: $\kappa = \frac{p_0 - p_c}{1 - p_c}$ where p_0 is the agreement between the predictions and p_c is the agreement with respect to a random prediction	**(Cohen's) Weighted Kappa:** (Cohens's) Kappa κ with additional weights included, e.g. according to risks or costs

See also (10) and (24) for a more detailed overview of such metrics. In Tab. 1, only the F_β score and the weighted (Cohen's) Kappa allow the integration of additional weights. For the F_β score, the factor β determines the relation of weights between precision and sensitivity (recall). For the weighted (Cohen's) Kappa, the weights can be more directly utilized to integrate risk factors. (24) All other metrics only depend on the TP, FP, TN, and FN values, directly or indirectly. Within the literature study, all of these metrics (and diagrams) were collected and documented, independent of whether they had been applied in the training, validation, and/or testing phase.

The overall rate of publications, which included risk factors was addressed as the primary endpoint. No formal hypothesis testing and a-priori estimation of statistical power was included. But, an a-posteriori estimation (one-sided 95% confidence interval) for the inclusion of risk factors was performed assuming a binomial distribution. For this purpose, the **binom.test** function from the R statistical computing package (version 4.0.5, The R Foundation for Statistical Computing, Vienna/Austria) was used. This function applies the Clopper-Pearson interval for the estimation of the confidence interval.

Remark: The term validation in this paper refers to the fine tuning of ML models / selection of hyperparameters, as it is commonly used in the ML community. In classical terms regarding development processes, validation means "… establishing by objective evidence that device specifications conform with user needs and intended use(s)" (25). In this sense, validation does not only refer to a tuning of models using independent data but to a proof that the technical criteria meet the needs of the particular application. Thus, not only technically sound performance metrics should be used, which are based on the number (like Acc, $F1$, or MCC), but their actual impact in the given use scenario need to be considered. Otherwise, this more general notion of validation cannot be addressed, appropriately.

2.2 Topic B – Impact of risk factors into performance metrics

As a second topic, the impact of risk factors was assessed, when they are integrated into performance measures for binary classification tasks. For this purpose, an artificially constructed model was utilized for the error distributions as well as a modification of the accuracy measure, in this paper. The model was first introduced in (23). It includes dedicated weight factors which represent the costs of the different types of errors. This reflects a limited version of the full decision theoretic approach as proposed in (16, 21). Instead, it was more directly adjusted towards its use in ML-based classification tasks. In particular, the model was coupled to the corresponding ROC curves, for this purpose. In comparison to references like (16, 21, 22), we utilized a different notation which does not require the full background about decision theory and utility functions, but provides a self-explanatory description. Basically, the model implements a single level of risk factors. Deeper hierarchies of influencing parameters, like cascaded probabilities, further uncertainty factors, or value-of-information aspects, were not included (21). Additionally, the rational / normative approach of decision theory was pursued, as initially proposed by von Neumann and Morgenstern (26). This focuses on a purely probabilistic modelling and linear weights with respect to risk factors, i.e. the utility function is a sum of the severities of harm multiplied by their likelihoods. Aspects like non-linear utility functions, e.g. for implementing risk aversion or risk seeking policies (16), were not addressed.

In this paper, the following artificially constructed model for the performance of the classifier was applied to get better control of the classifier's behavior. A generic setup was used with a classifier F predicting the binary outcome $Y \in \{0,1\}$ from a set of input features X, i.e. the prediction is performed according to $\hat{Y} = F(X)$. This prediction was considered to be applied to a set of data (X_i, Y_i), where the Y_i were considered as the ground truth, i.e. the correct classification values for the input values X_i. The (X_i, Y_i) could represent training, validation, or test data. Additionally, it was regarded that the classifier depends on a threshold s. Thus, a particular instance of the classifier can be represented by a binary-valued function $F(s, X)$ which includes the threshold s as a parameter. As already mentioned, we utilized an artificially constructed error distribution to demonstrate the behavior of performance metrics when certain parameters get changed. This means, that we assumed that the false positive $FPR(s)$ and false negative rates $FNR(s)$ are given by a parametric function. We used modified Gaussian functions of the following form, for this purpose.

$$FPR(s) = (1 - s) \cdot \exp\left(-\frac{s^2}{\sigma_{FP}}\right) \qquad (1)$$

$$FNR(s) = s \cdot \exp\left(-\frac{(1 - s)^2}{\sigma_{FN}}\right) \qquad (2)$$

The included terms $(1-s)$ and s modify the Gaussians in a way that $FPR(1) = FNR(0) = 0$. Fig. 1, left side shows the course of the error distributions along the threshold s and for the parameter set $\sigma_{FP} = \sigma_{FN} = 0.3$. On the right side, the corresponding ROC curves are shown for varying parameters. Mind that the threshold s is only encoded implicitly, in the ROC curve representation.

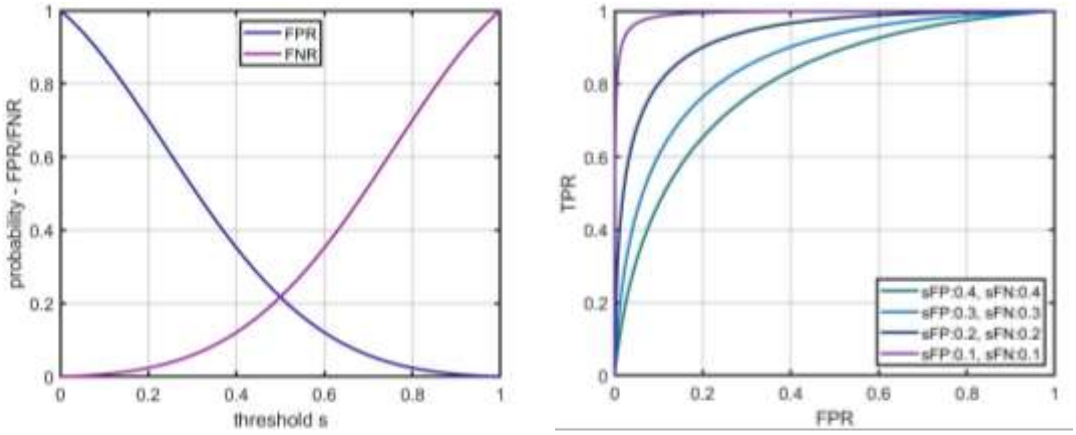

Fig. 1. Left side: Artificial model of error distributions, i.e. $FPR(s)$ and $FNR(s)$ in dependence of the threshold s. The model is based on the modified Gaussian functions as defined in equations (1) and (2), i.e. of the form $FPR(s) = (1-s) \cdot \exp\left(\frac{s^2}{\sigma_{FP}}\right)$ and $FNR(s) = s \cdot \exp\left(\frac{(1-s)^2}{\sigma_{FN}}\right)$. Left side: model with fixed parameters $\sigma_{FP} = \sigma_{FN} = 0.3$. Right side: Resulting ROC curves for a set of different parameters, $\sigma_{FP} = \sigma_{FN} = 0.1$, $\sigma_{FP} = \sigma_{FN} = 0.2$, $\sigma_{FP} = \sigma_{FN} = 0.3$ and $\sigma_{FP} = \sigma_{FN} = 0.4$.

As a next step, a risk model was constructed which assigns certain “costs” to the different types of errors FP and FN. These costs reflect the impact of the particular risks which are caused by the corresponding type of error. We assume costs w_{FP} and w_{FN}, which are fixed weights. In the current paper, we do assume no costs for the cases of correct classifications, but only for the error cases. In terms of conditional probabilities $P(\hat{Y}|Y)$, the resulting expected risk $ER(s)$ can be calculated according to

$$ER(s) = E\left(w_{FP} \cdot P(\hat{Y} = 1|Y = 0) + w_{FN} \cdot P(\hat{Y} = 0|Y = 1)\right), \qquad (3)$$

where $E(\cdot)$ denotes the expected value. For given numbers of positive and negative cases, i.e. P and N, the expected risk can be calculated as

$$ER(s) = w_{FP} \cdot N \cdot FPR(s) + w_{FN} \cdot P \cdot FNR(s). \tag{4}$$

Positive and negative refers to the true situation, i.e. true prevalence, and not the predictions, since only these relationships reflect the actual use case. Basically, the expected risk $ER(s)$ can be considered as a negative version of a utility function, since it represents some kind of costs instead of utilities / benefits. This is consistent with the general definition in normative decision theory (22), where the expected utility $EU(s)$ is defined as the sum of utilities $U(r)$ across all potential outcomes r from a set R of results weighted by the respective probabilities $P(\text{Result}(s) = r|s)$, i.e.

$$EU(s) = \sum_{r \in R} U(r) \cdot P(\text{Result}(s) = r|s). \tag{5}$$

$P(\text{Result}(s) = r|s)$ represents the probability, that the outcome r occurs, when a given parameter or threshold s is used. In general, the formula can be conditioned with respect to an additional evidence e (22). But, this was not further pursued in our paper. For the results set $R = \{FP, FN\}$, we obtain the relationships $U(FP) = w_{FP}$, $U(FN) = w_{FN}$, $P(\text{Result}(s) = FP|s) = P(\hat{Y} = 1|Y = 0) = N \cdot FPR(s)$, and $P(\text{Result}(s) = FN|s) = P(\hat{Y} = 0|Y = 1) = P \cdot FNR(s)$. This represents the basic relationship between our approach and normative decision theory. Mind that in our case, we used costs instead of utilities. This clarifies in which way the expected risk $ER(s)$ represents a negative version of a utility function.

For finding the best threshold s, the expression $EU(s)$ has to be maximized respectively $ER(s)$ minimized. We can apply a monotone transformation on $ER(s)$ without changing the relationships between ER values and thus also the optimization procedure. In general, linear transformations do not substantially change a utility function (22). In particular, a linear transformation of the following form can be applied to obtain modified, but equivalent values $\widetilde{ER}(s)$:

$$\widetilde{ER}(s) = \frac{1}{w_{FP} \cdot N} ER(s) = FPR(s) + \frac{w_{FN} \cdot P}{w_{FP} \cdot N} \cdot FNR(s). \tag{6}$$

Using the relative proportion

$$c_{FN} = \frac{w_{FN} \cdot P}{w_{FP} \cdot N}, \qquad (7)$$

this modified version can be written in a simpler form as

$$\widetilde{ER}(s) = FPR(s) + c_{FN} \cdot FNR(s). \qquad (8)$$

Subsequently, c_{FN} is called risk ratio as it reflects the relationship between the error types FN and FP. Such a simplification, where only the relative ratio of risk values is considered, is limited to the case when only two risk factors are regarded. $\widetilde{ER}(s)$ will still be called expected risk since it is equivalent to $ER(s)$ with regard to risk minimization as given in the following formula. In other words, the formula determines the threshold s which optimizes the expected risk, i.e.

$$s = \underset{s}{\operatorname{argmax}}\, ER(s) = \underset{s}{\operatorname{argmax}}\, \widetilde{ER}(s) = \underset{s}{\operatorname{argmax}}\big(FPR(s) + c_{FN} \cdot FNR(s)\big). \qquad (9)$$

This turns the task of finding the threshold for the binary classification problem into a decision problem with respect to the expected risk. In contrast to many standard scenarios in decision theory, it is not a decision between a set of discrete alternatives or actions but between different values of the threshold s coming from a continuous range of alternatives. However, it remains the decision for a certain value under the uncertainties given by the particular risks. This procedure can be represented as shown on the left side of Fig. 2, where the expected risk $\widetilde{ER}(s)$ for the artificial model given by (1) and (2) is plotted against the threshold value. The optimum threshold is the point where the function $\widetilde{ER}(s)$ achieves its minimum. The position of the minimum is shown by the dotted line. Due to the symmetry of the artificial model, this line lies at $s = 0.5$.

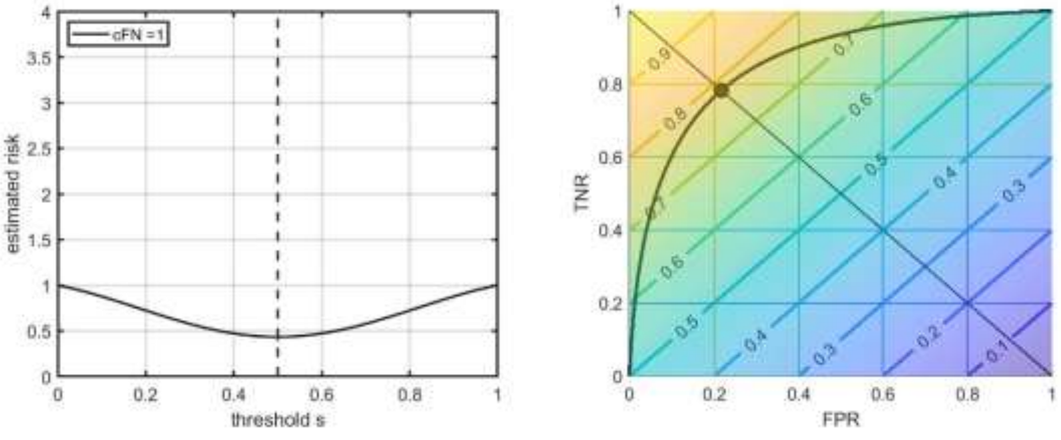

Fig. 2. Left side: Representation of the threshold optimization with respect to the expected risk $\widetilde{ER}$ using a diagram where the x axis represents the threshold variable s and the y axis the $\widetilde{ER}(s)$ function. The same artificial model was used as in Fig. 1, left side (i.e. with parameters $\sigma_{FP} = \sigma_{FN} = 0.3$). The optimum threshold is the point where $\widetilde{ER}(s)$ reaches its minimum. Right side: ROC diagram for the same model with the WBA metric overlaid in a color coding as well as its contour lines. The optimization of WBA is equivalent to finding the optimum threshold for the expected risk $\widetilde{ER}$**.** In the representation on the right side, (local) optimization of WBA is equivalent to finding the points on the ROC curves which are tangent to the iso-contour lines of the function WBA (depicted by the dot). The diagonal line represents the symmetry line between positive and negative cases.

The expected risk can be considered as a performance metric for classifiers which integrates a risk-based weighting to the error rates. In contrast to usual metrics, the lower values describe a better performance since errors are counted and not the rate of correct assignments. However, this can be converted into each other. For this purpose, we apply another linear transformation to obtain the following metric, which is subsequently called weighted balanced accuracy (WBA).

$$WBA(s) = \frac{1 + c_{FN} - \widetilde{ER}(s)}{1 + c_{FN}} = \frac{1 + c_{FN} - \big(FPR(s) + c_{FN} \cdot FNR(s)\big)}{1 + c_{FN}}$$

$$= \frac{\big(1 - FPR(s)\big) + c_{FN} \cdot \big(1 - FNR(s)\big)}{1 + c_{FN}} = \frac{TPR(s) + c_{FN} \cdot TNR(s)}{1 + c_{FN}}$$

$$= \frac{1}{1 + c_{FN}} \cdot TPR(s) + \frac{c_{FN}}{1 + c_{FN}} \cdot TNR(s) = w_{TP} \cdot TPR(s) + w_{TN} \cdot TNR(s). \qquad (10)$$

This shows, that $\widetilde{ER}(s)$ is indeed equivalent to a weighted version of the balanced accuracy metric $BA = \frac{FPR(s) + FNR(s)}{2}$, where $w_{TP} + w_{TN} = 1$, i.e. the weights add up to 1. This guarantees that the maximum value of this metric equals 1 as well. Due to the relationship $c_{FN} = \frac{w_{FN} \cdot P}{w_{FP} \cdot N}$, the weights are basically determined by the true prevalence, i.e. the relationship between actual positive and the total number of cases, as well as the relationships of the costs w_{FN}, w_{FP} between the particular types of errors. As long as the risk ratio c_{FN} equals 1, the expected risk is equivalent to the balanced accuracy BA. $c_{FN} = 1$ reflects the situations where the effects of prevalence and risk weighting balance out, i.e. when

$$w_{FP} \cdot N = w_{FN} \cdot P. \tag{11}$$

This relationship will be utilized later in section 3.3 when considering standard schemes for risk assessment.

A graphical representation of this weighted balanced accuracy metric WBA is shown on the right side of Fig. 2, in combination with the ROC curve. WBA is depicted using a color coding which represents the value of the function (yellow / light colors represent the highest values). Additionally, the iso-contour lines of this function are portrayed in order to make the course of the function better accessible. In this representation, optimization with respect to the threshold is the same as finding the points on the ROC curve which are tangent to the WBA or equivalently the $\widetilde{ER}$ function. More precisely, the tangents of the ROC curve need to be tangential to the iso-contour lines of WBA. Basically, this procedure achieves a local optimization. A selection of the tangent at the point with the highest WBA (or lowest $\widetilde{ER}$) value has to be performed in the case of multiple local optima. In the diagram, the optimum point of the ROC curve is shown as a dot. In this diagram, the symmetry is characterized by the diagonal line. Mind, that the threshold s is not encoded explicitly here. It is only given implicitly by the correspondence between the points on the ROC curve and the corresponding threshold values for the analyzed model.

As a next step, the impact of different risk ratios was analyzed for the model given in Fig. 1 respectively in equations (1) and (2) as an example to demonstrate the analysis method. For this purpose, it was assumed that the optimum threshold $s_{1.0}$ had been determined using an $\widetilde{ER}$ function without a risk-based weighting, i.e. when $c_{FN} = 1.0$. Basically, this leads to a metric which is equivalent to the balanced accuracy BA. Then, this threshold $s_{1.0}$ was applied to the $\widetilde{ER}$ function with a risk-based weighting included, i.e. $c_{FN} \neq 1$. In this example, $c_{FN} = 0.25$ and $c_{FN} = 4.0$ was used. The resulting value $\widetilde{ER}(s_{1.0})$ was compared to the situation where the thresholds $s_{0.25}$ and $s_{4.0}$ would have been used, i.e. to the situation, when the expected risk would have been obtained with the correct weight $c_{FN} \neq 1$. The effect of this variation is shown in Fig. 3. In the upper row, the $\widetilde{ER}(s)$ values were plotted against the threshold s. For comparing the results, the threshold $s_{1.0}$ (located at the midline $s = 0.5$) as well as the height of the expected risk at $s_{1.0}$ was included in the diagrams for $c_{FN} = 0.25$ and $c_{FN} = 4.0$ as dashed black lines. The optimum thresholds and corresponding expected risks are shown by the blue (for $c_{FN} = 0.25$) and turquoise (for $c_{FN} = 4.0$) line elements. The resulting difference between the risk values (at $s_{1.0}$ vs $s_{c_{FN}}$) is shown by the Δ symbol at the side.

In the bottom row of Fig. 3, the situation is shown using the ROC curves enriched with the WBA metric. The iso-contours remained straight lines but their slope changed according to the different weights of positive and negative cases. This had an impact on the determination of the optimum points, since the tangents between the ROC curve and the iso-contours now match at another position. These optimum points $s_{0.25}$, $s_{1.0}$, and $s_{4.0}$ in ROC space were depicted by black dots. It can be seen, that the optimum now deviates from the diagonal symmetry line. For the cases with $c_{FN} \neq 1$, the default threshold $s_{1.0}$, i.e. the threshold for the case $c_{FN} = 1$, is shown as a white dot.

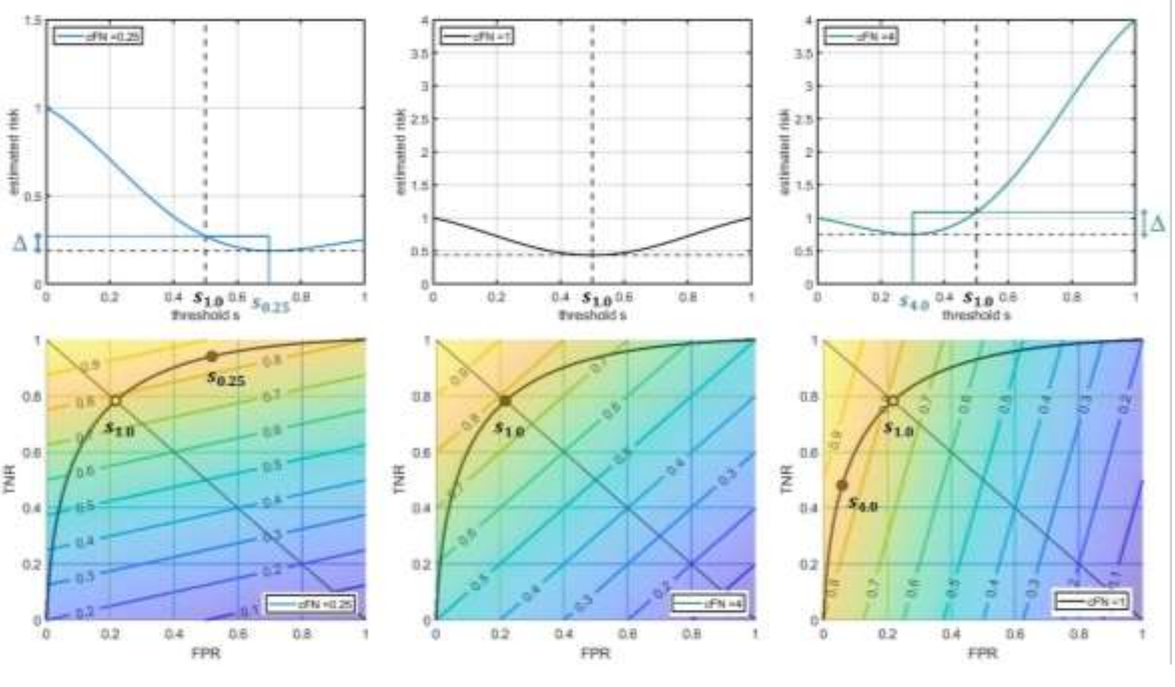

Fig. 3. Upper row: Impact of different risk ratios $c_{FN} = 0.25, 1.0,$ and 4.0 (from left to right) on the threshold selection and the resulting estimated risk $\widetilde{ER}(s)$, which is shown on the y axis. The same artificial error distribution was used as in Fig. 1. The default threshold $s = 0.5$ (for the case $c_{FN} = 1.0$) and the

corresponding estimated risk is depicted as the black dashed line in all three cases. The difference between this default and the true optimal threshold $s_{0.25}$ and $s_{4.0}$ is shown by the additional blue (for $c_{FN} = 0.25$) and turquoise (for $c_{FN} = 4.0$) lines. The resulting difference in the $\widetilde{ER}(s)$ values is marked by the symbol Δ. Mind that a different scaling of the y axis was used in the $c_{FN} = 0.25$ case in order to better visualize the differences. Bottom row: ROC curves for the same cases enriched with the WBA (weighted balanced accuracy) metric. A color coding and the corresponding contour lines are used to visualize the course of the function. The optimum points in ROC space for the particular risk ratios c_{FN} (again named $s_{0.25}$ and $s_{4.0}$) are given by the black dots. They represent the points where the tangent of the ROC curve and the iso-contour of the WBA metric coincide. The white dot refers to the default threshold $s = 0.5$ and makes the differences of the threshold estimation visible.

This describes the basic approach for our analysis. This was applied to a more comprehensive setting in order to systematically elaborate the effect of different risk ratios on the expected risk and the associated metrics. For this purpose, the risk ratio c_{FN} was systematically varied from $\frac{1}{16} = 2^{-4}$ to $16 = 2^4$. The increment for the risk ratios between the steps was given by a factor of 2. Additionally, the risk ratios $c_{FN} = 0.1$ and $c_{FN} = 10.0$ were included, since they represent important references with respect to the application of risk management in medical devices. This is demonstrated later in section 3.3. Further on, the parameters of the artificial model / error distribution, as given by the modified Gaussians (1) and (2), were varied. The parameter sets $\sigma_{FP} = \sigma_{FN} = 0.1$, $\sigma_{FP} = \sigma_{FN} = 0.2$, $\sigma_{FP} = \sigma_{FN} = 0.3$ and $\sigma_{FP} = \sigma_{FN} = 0.4$ were used. The overall relative difference in $\widetilde{ER}(s)$ values when applying these changes was the main endpoint of this part of the study. The implementation of the calculations was performed using Matlab (version R2021a, The MathWorks Inc., Natick/ Massachusetts).

2.3 Topic C – Integration into the development process for ML-based medical devices

Finally, an analysis of the regulatory requirements was performed which have to be fulfilled within the development of ML-based medical devices. In particular, the requirements on risk management and their relationship to the evaluation of ML-based classification models were addressed.

Basically, the analysis in this paper focused on the requirements in the European Union (EU). Thus, the Medical Device Regulation (MDR) (6) was considered as the central reference. Subsequently, the corresponding (harmonized) standards have to be respected as well. For risk management, this is ISO 14971 (8). Additionally, the technical report ISO/TR 24971 (9) was taken into account. It provides further guidance how to implement risk management into the development of medical devices. As a second upcoming regulation, the proposed AI Act of the EU (7) and its relevant requirements, e.g. regarding risk management, data governance, or quality management, were included.

Basically, the impact of these regulations and standards on the definition of appropriate performance metrics was analyzed, within this paper. In particular, the requirements for the inclusion of risk factors instead of purely applying standard metrics like Acc, $F1$, or MCC were examined. Additionally, the analysis elaborated challenges and potential improvements for a consequent risk-based approach towards the evaluation of ML-based classification models. This was addressed utilizing the following two main applications and use scenarios. For each application, a series of modifications was included to demonstrate the impact of different risk factors on assessment of model performance.

Use Scenarios

A. *diagnostic test:* ML-based system which is integrated into a screening test for a specific disease (e.g. a specific type of cancer). The actual prevalence of the disease as well as the probabilities of different types of errors / risks, i.e. TP, FN, TN, and FP, are assumed to be fixed in the following subcases.

   1. situation with very high risk in case of false negatives (FN), when an early detection of the disease is missed, e.g. because it quickly develops into a critical state where the success rate of potential treatments is very limited

2. situation still with high risk in case of false negatives (FN), because the impact of the disease basically is serious, but with an option to better detect the disease by additional tests
3. situation with reduced risk in case of false negatives (FN), because the disease develops rather slowly and has less severe impact
4. situation with reduced risk in case of false negatives (FN), like in scenario A3, but additionally with high risk in the case of false negatives (FP), e.g. when a biopsy or another treatment needs to be performed in the case of positively predicted cases (i.e. TP and FP), which may cause substantial harm to the patient

B. *quality inspection:* ML-based quality assurance system for identifying deficiencies in surgical instruments before they get delivered. It is assumed that the same ratio relationships between positive (instrument has a defect) and negative cases (instrument has no defect) as well as error cases (i.e. TP, FN, TN, and FP) is given as in use scenario A.

1. situation where instruments with a missed detection of a defect (FN) will be delivered directly to a hospital and may cause serious harm to a patient when applied in the treatment procedure
2. situation as in case B1, but this time including an additional check in the hospital which substantially lowers the probability and/or severity of the potential harm of FN cases
3. situation where the quality assurance step is designed to identify defects in an early production step and eliminate the particular instrument to reduce further financial costs, caused by FP. In this case, it is considered that additional quality steps are included to keep the FN rate at an appropriate level, e.g. additional visual inspections or tests, which reduce the risk of delivering defect instruments / producing harm on the patient to a low and acceptable level.

3 Results

3.1 Topic A – Utilization of risk-based performance metrics in recent publications

The literature search for analyzing how often risk-based approaches are used in current scientific literature was performed on Nov 15, 2022. According to the option "*in the last 1 year*", it included papers from Nov 2021 to Nov 2022. The analysis was done by the first author, based on the search strategy as described in section 2.1. For the given search term, 115 publications were found in total. Starting from the most recent publication, 55 papers were analyzed, since 25 of them had to be excluded according to the given criteria. These publications and the corresponding reasons for exclusion are provided in table S1 (supplements). Based on this, 30 papers were finally included, as defined in the search strategy. These publications were analyzed in detail. The performance metrics, used for binary classification tasks in the particular publications are listed in Tab. 2. In some cases, additional metrics were included which we did not have on our initial list. They were also documented in Tab. 2. None of them included risk factors, in a dedicated way.

Tab. 2. Analysis of articles which were included in the literature research regarding recent publications about performance metrics of ML-based classification models (sorted according to the "most recent" criterion). The table documents the used performance metric as well as the rating regarding the inclusion of risk-based elements.

first author + ref no.	used performance metrics	inclusion of risk-based elements
Ozcan (27)	Acc, Sen, Prec Additional metrics (without direct risk integration): Determinism → was neither described nor referenced reliably	No
Garavand (28)	Acc, Prec, Sens, Spec, F1 Score, ROC, AUROC, AUPRC	No
ElSeddawy (29)	Acc, Sens, Spec, F1 Score, G-mean, ROC, AUROC, (unweighted) Kappa	No

Kasim (30)	Acc, Prec, NPV, Sen, Spec, AUROC, (unweighted) Kappa Additional metrics (without direct risk integration): net reclassification index (NRI)	In this case, the basic application (mortality prediction) was strongly related to a risk-based application itself. Thus, also the evaluation included risk factors, in some sense, even though standardized metrics were used. The effect, which were caused by errors in the ML systems itself, were not included additionally.
Aldhyani (31)	Acc, Prec, Sen, Spec, F1-score	No
Wu (32)	Acc, Prec, Sen, F1-Score, ROC, AUROC	No
Preto (33)	Acc, Prec, Sen, F1-Score, AUROC	No
González-Cebrián (34)	Acc, Sen, Spec, F1-Score, MCC, AUROC	In this case, the basic application (mortality prediction) was strongly related to a risk-based application itself. Thus, also the evaluation included risk factors, in some sense, even though standardized metrics were used. The effect, which were caused by errors in the ML systems itself, were not included additionally.
He (35)	Acc, Prec, Sen, F1-Score, ROC, AUROC	No
Milara (36)	Acc, Prec, Sen, Spec, F1-Score, AUROC	No
Emakhu (37)	Acc, Prec, Sen, Spec, MCC, F1 score, ROC, AUROC	In this case, the basic application (Acute coronary syndrome prediction) was related to a risk-based application itself. Additionally, there was a cost-sensitive approach included in the evaluation of the models, besides the utilization of standardized metrics.
Haq (38)	Acc, Prec, NPV, Sen, Spec, ROC, Additional metrics (without direct risk integration): Dice Similarity Coefficient (DSC), Probabilistic Random Index (PRI).	No
Movahed (39)	Acc, Sen, Spec, F1-Score, ROC, AUROC Additional metrics (without direct risk integration): False Discovery Rate	No
Templeton (40)	Acc, Prec, Sen	No
Zou (41)	Acc, BA, Prec, Sen, Spec, F1-Score, MCC, ROC, AUROC	No
Tran (42)	Acc, F1-Score, ROC, AUROC	No
Maskew (43)	Acc, PPV, NPV, ROC, AUROC	No

Mabrouk (44)	Acc, BA, Prec, Sens, F1 score	No
Khan (45)	Acc, Prec, Sens, F1 score	No
Ho (46)	Acc, Prec, Sens, F1 score	No
Eissa (47)	Acc, Prec, Sens, MCC, F1 Score, ROC, AUROC	No
Salimpour (48)	Acc, Prec, Sens, (unweighted) Kappa	No
Berenguer-Vidal (49)	Acc, Prec, Sen, Spec	No
Dritsas (50)	Acc, Prec, Sens, F1 Score, AUROC	No
Ahmad (51)	Acc, Prec, Sen, Spec, ROC	No
Goñi (52)	BA, Prec, NPV, Sens, Spec, ROC, AUROC	No
Dubol (53)	Acc, AUROC	No
Hidayat (54)	Acc, Sen, Spec, ROC, AUROC	No
Baskozos (55)	BA, MCC, AUPRC	No
Shakhovska (56)	Acc, Prec, Sens, F1 Score, AUROC	No

In total, only 3 out of the 30 publications, i.e. the papers (30), (34), and (37), included a risk-based approach for the performance assessment of the ML models, in some sense. Basically, all of these three publications were addressing risk prediction as the major application. Thus, they had the risk assessment part integrated according to the direct nature of the application. In two cases, i.e. (30) and (34), the ML models were developed for mortality prediction. The concrete use of the ML models in clinical practice as well as the potential impact of errors was not addressed and not included in the evaluation, in these cases. In (37), the main goal of the development was the prediction of an acute coronary syndrome. Additionally, a cost-sensitive approach was included in the evaluation of the models, besides the utilization of standardized metrics. This was the only case, where risk- or cost-based elements were included in the evaluation, directly. For all other cases, only standardized metrics were included. Neither the F_β score nor the weighted (Cohen's) Kappa was used, which would basically allow to integrate risks or costs as weight factors.

Based on these results, there were different alternatives, how to count these cases. For this reason, we included the following three different estimations for the one-sided 95% confidence interval (CI). In any case, the CI was calculated as a Clopper-Pearson interval as defined in 2.1.

- Case AI: The three publications (30), (34), and (37) (out of a total of 30 publications), which had some kind of risk prediction, were considered as positive results. In this case, there was a 10% rate (3 out of 30) of publications including risk factors. The upper limit of the 95% CI was 0.24, i.e. 24%.
- Case BII: The two cases (30) and (34), which addressed mortality prediction as the application of the ML model and which did not include any further risk-based elements in the evaluation of the models, were excluded. The paper (37), which included risk factors in the evaluation , were counted as the only remaining positive case. This led to an overall result of 1 in 28 cases, i.e. a 3.6% rate. Here, the 95% CI was 0.16, i.e. 16%.
- Case CIII: All cases, where a risk prediction was the main objective of the model itself, were excluded. Thus, there were 0 positive out of 27 total case, leading to a 0% rate and an upper limit of the 95% CI of 0.11, i.e. 11%.

3.2 Topic B – Impact of risk factors into performance metrics

This section demonstrates how changes in the risk factors affect the evaluation of ML classification models. For this purpose, Tab. 3 and Fig. 4 show the results of the expected risk $\widetilde{ER}$ which were obtained, when varying the risk ratio c_{FN} systematically between $\frac{1}{16} = 2^{-4}$ to $16 = 2^4$, with an increment by factor 2 between the steps. Additionally, the impact for the values $c_{FN} = 0.1$ and $c_{FN} = 10.0$ was evaluated. For visualization purposes, the range for c_{FN} was reduced to $\frac{1}{8} = 2^{-3}$ to $8 = 2^3$ in the left part of Fig. 4. For the evaluation, the artificial model given in (1) and (2) was used where the parameter for the modified Gaussians were set to $\sigma_{FP} = \sigma_{FN} = 0.1$, $\sigma_{FP} = \sigma_{FN} = 0.2$, $\sigma_{FP} = \sigma_{FN} = 0.3$, and $\sigma_{FP} = \sigma_{FN} = 0.4$. The expected risk values given at the default threshold $s_{1.0} = 0.5$ were compared to the outcome at the optimum threshold $s_{c_{FN}}$ for the particular risk ratio c_{FN}.

The main results are provided in the right most column of Tab. 3, in terms of the relative difference between $\widetilde{ER}\left(s_{c_{FN}}\right)$ and $\widetilde{ER}(s_{1.0})$. It can be seen that this relation goes up to 2.98, i.e. 198% increase in expected risk, for the parameter setting $\sigma_{FP} = \sigma_{FN} = 0.4$ and the risk ratio $c_{FN} = 10.0$. For $c_{FN} = 16.0$, this further increases to a relative difference of 4.55, i.e. an increase of 355%. The effect is less intense when the risk ratio is closer to $c_{FN} = 1.0$, i.e. the non-weighted case. For example, the increase is less than 12% for a risk ratio $c_{FN} \leq 2.0$. The described effects were also reduced in a certain degree when the values σ_{FP}, σ_{FN} decreased. Such a decrease implies that the ROC curve lies closer to an ideal model, as it can be seen in Fig. 1 right side.

Tab. 3. Differences of expected risk $\widetilde{ER}$ when varying the risk ratio c_{FN} systematically between 1.0 to $16 = 2^4$ (stepwise increment by factor 2) as well as $c_{FN} = 10.0$ as an extra point of evaluation. Due to symmetry reasons, the values for $c_{FN} < 1.0$ are equivalent to the inverse risk ratio $\frac{1}{c_{FN}}$. The rightmost column shows the relative differences between $\widetilde{ER}\left(s_{c_{FN}}\right)$, i.e. the value at the optimum position $s_{c_{FN}}$ for the particular curve, and $\widetilde{ER}(s_{1.0})$, i.e. the value at the default threshold $s_{1.0}$.

parameter settings of artificial model / risk ratio		**optimum threshold $s_{c_{FN}}$ and corresponding $\widetilde{ER}$ value**			**comparison of $\widetilde{ER}$ values: $s_{c_{FN}}$ vs default threshold $s_{1.0}$**
modified Gaussian σ_{FP} / σ_{FN}	risk ratio / weight c_{FN} / c	optimum threshold $s_{c_{FN}}$	estimated risk value		relative difference
			at $s_{c_{FN}}$: $\widetilde{ER}\left(s_{c_{FN}}\right)$	at $s_{1.0}$: $\widetilde{ER}(s_{1.0})$	$\frac{\widetilde{ER}(s_{1.0})}{\widetilde{ER}\left(s_{c_{FN}}\right)}$
$\sigma_{FP} = 0.1$, $\sigma_{FN} = 0.1$	1.0 (default)	0.5 (default)	0.08	0.08	1.0
	2.0	0.46	0.11	0.12	1.07
	4.0	0.44	0.16	0.21	1.30
	8.0	0.40	0.21	0.37	1.77
	10.0 (one level up)	**0.38**	**0.23**	**0.45**	**1.98**
	16.0	0.36	0.27	0.70	2.58

$\sigma_{FP} = 0.2$, $\sigma_{FN} = 0.2$	1.0 (default)	0.5 (default)	0.29	0.29	1.0
	2.0	0.44	0.40	0.43	1.08
	4.0	0.36	0.52	0.72	1.38
	8.0	0.3	0.65	1.29	1.97
	10.0 (one level up)	**0.26**	**0.70**	**1.58**	**2.26**
	16.0	0.22	0.78	2.44	3.12
$\sigma_{FP} = 0.3$, $\sigma_{FN} = 0.3$	1.0 (default)	0.5 (default)	0.43	0.43	1.0
	2.0	0.4	0.59	0.65	1.10
	4.0	0.3	0.75	1.09	1.44
	8.0	0.18	0.89	1.96	2.20
	10.0 (one level up)	**0.16**	**0.92**	**2.39**	**2.59**
	16.0	0.08	0.98	3.69	3.78
$\sigma_{FP} = 0.4$, $\sigma_{FN} = 0.4$	1.0 (default)	0.5 (default)	0.54	0.54	1.0
	2.0	0.36	0.72	0.80	1.11
	4.0	0.22	0.88	1.34	1.51
	8.0	0.08	0.98	2.41	2.45
	10.0 (one level up)	**0.04**	**1.00**	**2.94**	**2.96**
	16.0	0.00	1.00	4.55	4.55

The results are shown graphically in Fig. 4 on the right side, using a logarithmic scaling of the x axis, i.e. for the risk ratio c_{FN}. The reference values $c_{FN} = 0.1$ and $c_{FN} = 10.0$ are indicated by a vertical red line. It can be recognized, that the relative difference between $\widetilde{ER}\left(s_{c_{FN}}\right)$ and $\widetilde{ER}(s_{1.0})$ is symmetric to the axis $c_{FN} = 1.0$ (or equivalently $\log_2 c_{FN} = 0$). This is due to the construction of the model which has a symmetry between the positive and negative cases. Basically, this means that the relative difference in expected risk is the same between a risk ratio c_{FN} and its inverse $\frac{1}{c_{FN}}$. Because of this equality, the c_{FN} values below 1 were omitted in Tab. 3. On the left side of Fig. 4, the actual expected risk values $\widetilde{ER}(s)$ are shown in a similar way as in Fig. 3, upper row. In this case, the different risk ratios between $\frac{1}{8} = 2^{-3}$ and $8 = 2^3$ are integrated into one diagram. Again, the default threshold $s_{1.0} = 0.5$ was marked by the dashed line. The optimum thresholds $s_{c_{FN}}$ for the other risk ratios are lying at the minima of the particular $\widetilde{ER}$ curves. They are depicted by the vertical small dashes. Thus, the relationship between $\widetilde{ER}(s_{c_{FN}})$ and $\widetilde{ER}(s_{1.0})$ can be recognized as the difference of the particular curve with respect to its height, when comparing the minima with the position where the dashed line and the curve intersect.

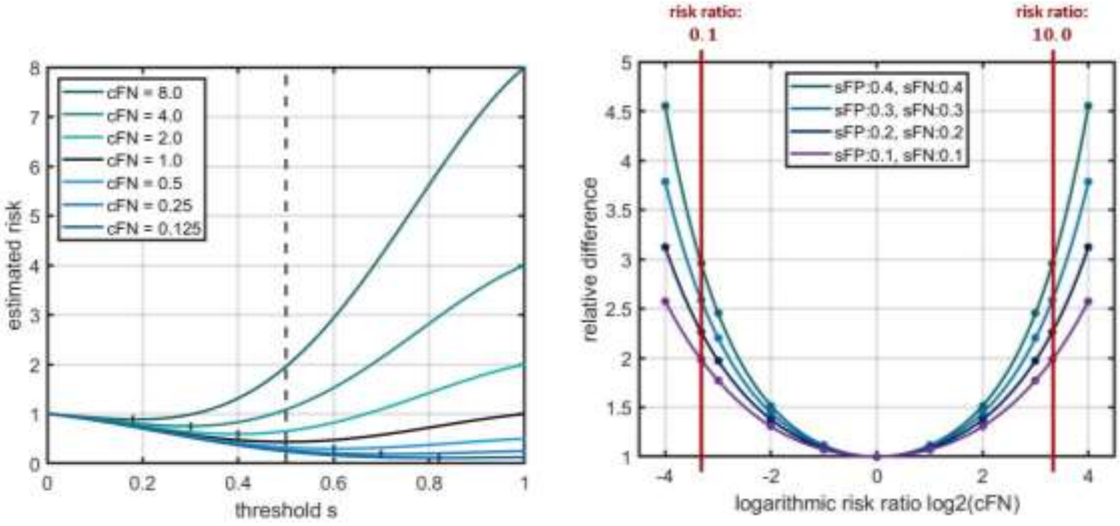

Fig. 4. Graphical representation of the results given in Tab. 3. Left side: Visualization of the expected risk ($\widetilde{ER}$) values for the particular risk ratios c_{FN} in the range $\frac{1}{8} = 2^{-3}$ to $8 = 2^3$ integrated into one diagram. The same artificial model as given in (1) and (2) was used. In this case, the model parameters were set to $\sigma_{FP} = \sigma_{FN} = 0.3$. The position of the default threshold $s_{1.0} = 0.5$ was marked by the dashed line. The optimum thresholds $s_{c_{FN}}$ for the other risk ratios were depicted by the small dashes (positioned at the minima of the particular $\widetilde{ER}$ curves). The intersection between the dashed line and the particular curve shows the $\widetilde{ER}(s_{1.0})$ value which can be compared to the minimum value, i.e. the optimum expected risk $\widetilde{ER}(s_{c_{FN}})$. Right side: Relative difference $\frac{\widetilde{ER}(s_{1.0})}{\widetilde{ER}(s_{c_{FN}})}$ across all risk ratios c_{FN}, i.e. the values in the right most column of Tab. 3, where a logarithmic scaling ($\log_2 c_{FN}$) was used on the x axis. The red lines mark the risk ratios $c_{FN} = 0.1$ and $c_{FN} = 10.0$, which typically represent a shift of one level in the risk matrix as described in section 3.3. Based on this.

the course of the relationship for different parameter settings of the artificial model ($\sigma_{FP} = \sigma_{FN} = 0.2$, $\sigma_{FP} = \sigma_{FN} = 0.2$, $\sigma_{FP} = \sigma_{FN} = 0.3$, and $\sigma_{FP} = \sigma_{FN} = 0.4$) can be identified.

3.3 Topic C – Integration into the development process for ML-based medical devices

Based on the results of the sections before, the relation of risk-based approaches for the evaluation of ML-based medical devices in comparison to the corresponding regulatory requirements was addressed. The analysis was focused on the requirements in the EU, as given in the MDR (6), the ISO 14971 (8) as the relevant standard for risk management, the ISO/TR 24971 (9) as a practical guidance for implementing risk management, and the proposed AI Act (7) as the future horizontal regulation for AI-based systems in the EU. According to Art. 6 in combination with Annex II of [7], ML-based medical devices typically will be assigned to the high-risk class of AI systems according to the proposed AI Act. In particular, this is the case for medical devices which have a potentially serious impact on the health of the patient, like in use scenario A (*diagnostic test*) of section 2.3. For such devices, a third-party, e.g. notified body, needs to be included into the conformity assessment, according to the MDR (6). This necessity is one of the guiding principles for the definition of high-risk AI systems in (7).

A similar classification applies to use scenario B (*quality inspection*) of section 2.3. In this case, the ML-based system is not directly included in a medical device, but represents a part of its production system. According to [7], the system is still considered a high-risk AI system as long as it represents a safety critical component of a medical device, which itself would be rated high-risk. Additionally, the ISO 13485 (57) as the standard for quality management systems requires that tools used in the production system need to undergo a computer system validation (CSV), if they potentially lead to risks in the application of the medical device. Thus, the evaluation of the ML-based models in the use scenario A and B should be addressed in a similar way.

Finally, the evaluation of medical devices and their components has to be related to clinical performance. This is a key aspect for the development of medical devices as required in the corresponding regulations, in particular in the MDR (6). Risks to the health of the patient have to be considered, since they constitute important clinical effects. According to (6), the risks, including single risks as well as the overall risk, have to be reduced as much as reasonably possible (ALARP principle). This has to be performed unless no further substantial improvement of the risk-benefit relation can be achieved. (6) This implies that the training, validation, and testing of ML-based models should include adjustments with respect to risk-based factors. Otherwise, the reduction of risks remains limited. Consequently, this limitation also applies to situations, where risk factors are only included during the adjustments / optimization of thresholds. Finally, a positive risk-benefit relationship has to be guaranteed. This potentially requires to include the positive impact of properly treated cases as well. This was omitted in the present paper, as we only focused on the risk factors. However, this can easily be integrated when considering benefits as negative versions of risk factors. The evaluation should reflect the concrete use case as given in the intended use of the medical device. Risk management needs to be performed in order to mitigate risk factors in exactly this direction, where the associated application context and user / patient population as well as normal use conditions, including foreseeable misuse, have to be regarded (8).

Within the development phase, state-of-the-art techniques in the particular domain have to be applied. For ML-based devices, this means that training, validation, and testing of the models has to be implemented according to appropriate and established performance metrics. This is also reflected in the proposed AI Act of the EU (7), which includes such requirements, e.g. in its articles about risk management (Art. 9), data governance (Art. 10), and quality management (Art. 17). In Art. 9, it is mentioned that "… testing shall be made against preliminarily defined metrics and probabilistic thresholds that are appropriate to the intended purpose of the high-risk AI system" (7). Additionally, "training, validation and testing data sets shall take into account, to the extent required by the intended purpose, the characteristics or elements that are particular to the specific geographical, behavioral or functional setting within which the high-risk AI system is intended to be used." (Art. 10 in (7)). Thus, it is important to consider the actual prevalence of the use case within the development and evaluation of an ML-based medical device.

Thus, the intended population should be addressed properly in the training, validation, and testing steps, when considering ML-based technologies. In the case of a classification task, e.g. for a disease or other deficiency, the intended population basically reflects the actual prevalence, i.e. the relative amount of positive case numbers. Thus, this number should be taken into account as a basic reference when developing an ML-based medical device. Currently, a balanced situation between positive and negative cases is often pursued for training, testing, and validation (11). This makes sense in order to balance the unreliability in the different groups and to address the requirement for fairness / non-discrimination as e.g. included in (7). In particular, this is important when the prevalence is a low number, e.g. the amount of positive cases lies in the order of 10^{-3} or lower. Such a situation is given in many situations. Usually, there are much more negatives than positives in the population, since the appearance of a disease or other deficiency often is limited unless an epidemic situation occurs. The reliability of ML-based models would be rather poor, if this ratio would be represented in the corresponding data sets. Thus, it makes sense to balance them by using a higher rate of positive cases than actually given. However, the final evaluation should reflect the actual prevalence according to the requirements described above.

For achieving this balance, the impact / costs of different types of errors need to be considered as well. With respect to risk management, the costs are related to the severity of the (potential) harm. This has to be multiplied with the probabilities to achieve an overall estimation of risks. In a certain sense, this is reflected by equations (7), i.e. $c_{FN} = \frac{w_{FN} \cdot P}{w_{FP} \cdot N}$, which characterizes the risk ratio as a combination of a ratio $\frac{w_{FN}}{w_{FP}}$ representing the costs and the ratio between negative and positive cases, which is related to the actual prevalence. A balanced situation occurs when the different effects are balanced out as given in equation (11), i.e. when $w_{FP} \cdot N = w_{FN} \cdot P$. This means that the relation between negative and positive cases respectively FP and FN needs to be reciprocal to the cost ratio to keep the overall risk ratio at a constant level. This relationship is shown graphically in Fig. 5 for different overall risk ratios c_{FN} between 0.125 and 8.0 with stepwise increment by factor 2.

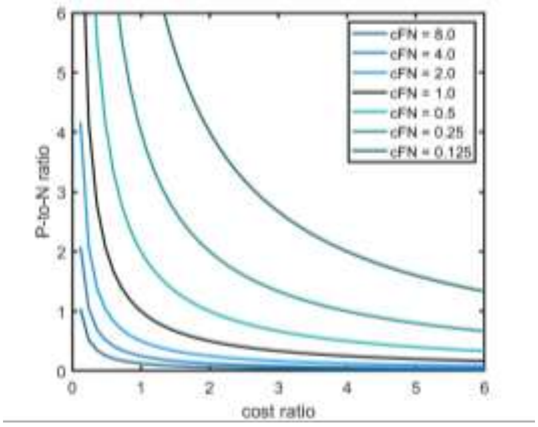

Fig. 5. Reciprocal relationship for the overall risk ratios c_{FN} (ranging from 0.125 and 8.0 with stepwise increment by factor 2). The product between the cost ratio $\frac{w_{FN}}{w_{FP}}$ for the particular risk and the relationship in numbers / probabilities needs to be constant to keep the overall risk at the same level.

The definition of risk as a combination of severity and probability is a central point in the risk management standard (8) and the associated guidance (9). In general, risk is considered as a situation that may lead to a harmful effect onto humans in some way, e.g. in terms of a physical harm. It is represented by a probability that this harm occurs and a severity which rates the level of impact. Ideally, this would be given in quantitative terms, i.e. concrete numbers for the probabilities and severities. However, it is recognized that this is often not possible in such a consequent way. Instead, it is allowed to perform risk analysis in a semi-quantitative or also qualitative way (8, 9). The semi-quantitative approach means that the probabilities and severities of risks are grouped together in certain levels, according to a rating performed by subject experts. The rating of the severities usually is done without giving concrete numbers, i.e. in a basically qualitative fashion. (8, 9) A typical example is the classification shown in Tab. 4 (see (9)):

Tab. 4. Semi-quantitative (with respect to probability levels) respectively qualitative (with respect to severities) classification of risks in medical devices as proposed in (9).

probability levels	**severity levels**
frequent: $\geq \mathbf{10^{-3}}$	negligible
probable: $< \mathbf{10^{-3}}$ and $\geq \mathbf{10^{-4}}$	minor

occasional: $< \mathbf{10^{-4}}$ and $\geq \mathbf{10^{-5}}$	serious / major
remote: $< \mathbf{10^{-5}}$ and $\geq \mathbf{10^{-6}}$	critical
improbable: $< \mathbf{10^{-3}}$	catastrophic / fatal

These categories basically reflect the probabilities which occur due to certain types of errors as given by the FPR and FNR values (for probabilities) as well as the particular 'costs' of errors respectively risk scores w_{FP} and w_{FN}. Usually, the probability levels are given with an exponential increase between these levels, e.g. in exponential steps with respect to the power 10, i.e. in levels of type 10^{-x}. The definition in Tab. 4 uses such an approach.

The relevant risks for a medical device are collected in a risk matrix as shown in Tab. 5. In this matrix, the particular risks are arranged in each combination of probability and severity levels. There typically are the following three areas contained in this matrix, which represent different requirements for further treatment of risks. (9)

- a red/orange area, where risks are considered as inacceptable and mandatorily need to be reduced before the medical device can be placed on the market – e.g. R_6 in Tab. 5
- a green area, where the risks can be regarded as insignificant and no further reduction needs to be considered – e.g. R_1, R_3, R_4 in Tab. 5
- a yellow area, sometimes called ALARP region, where risks need further investigation – e.g. R_2, R_5 in Tab. 5

The concrete ranges for the areas have to be prespecified in a risk policy, i.e. in the initial phase of the development within the risk management plan for the device (8, 9). Thus, acceptability of risks has to be assessed according to a strategy which is defined in advance.

Tab. 5. Risk matrix based on the risk semi-quantitative / qualitative classification as given in Tab. 4. The risk matrix collects all particular risks of a medical device (R_1 – R_6 in this case) according to its categorization with respect to their probability and severity (basic scheme as presented in (9)). The tree different areas (red/orange – inacceptable risks, green – acceptable risks, and yellow – region where risks need further investigation) indicate which further risk management steps have to be considered before the medical device can be placed on the market.

		severity levels				
		negligible	minor	serious / major	critical	catastrophic / fatal
probability levels	frequent	🟧	🟧	🟧	🟧	🟧
	probable	🟨	🟧	🟧	🟧	🟧
	occasional	🟩 $\boldsymbol{R_1}$	🟨	🟧	🟧	🟧
	remote	🟩 $\boldsymbol{R_3}$	🟩	🟨 $\boldsymbol{R_5}$	🟧 $\boldsymbol{R_6}$	🟧
	improbable	🟩	🟩 $\boldsymbol{R_4}$	🟩	🟨 $\boldsymbol{R_2}$	🟧

As already mentioned, the risks need to be considered as a combination between probabilities and severities. One standard approach is to calculate them by a multiplication between these two factors. (58) Other combinations may also be possible since (8, 9) do not specify further details about the combination. However, the multiplicative approach is consistent with the probabilistic method provided in section 2.2 as well as the normative version of decision theory. This approach is subsequently used to demonstrate the impact of different risk factors. In order to get a constant overall risk ratio, the probabilities need to be balanced with the associated severity level, i.e. their product needs to be equal to 1, in the multiplicative approach. For example, this can be applied to a situation where balanced data sets are used in combination with a standard performance metric, i.e. without additional weighting. In this case, a complete balancing between cost and probability ratios is implicitly assumed, i.e. the product between the severity and the probability ratio for the different types of errors is considered to equal 1.

The contributions of the different risk factors, e.g. $R_1 - R_6$ in Tab. 5, are usually considered to be additive. This means that the overall risk is a sum of the particular combined risks, in accordance with the formulas for expected risk presented in section 2.2. For example, the risks, i.e. the products of probabilities and severities / costs, can be summed up into a single weight, when one risk, e.g. one type of error, shows up with multiple severity and probability levels. The same applies to a situation, where multiple aspects need to be integrated into one particular type of risk. Thus, these situations are covered by the given approach. In general, there may be a more complex combination of several effects which go beyond the scope of this paper. Within this paper, we focused on only two particular risks, namely the risk for FN as well as the risk for FP. In this case, only the ratio $c_{FN} = \frac{w_{FN} \cdot P}{w_{FP} \cdot N}$ between them is relevant, when considering an ML-based classification task. Here, the values $w_{FN} \cdot P$ and $w_{FP} \cdot N$ aggregate the risks, i.e. severity times probability, for the particular type of error.

Typically, the elements at the diagonal of the risk matrix represent approximately constant levels of risk. If the probability levels are represented by an exponential scale with base 10, the severity levels also need to provide such increments in order to achieve this. Thus, we assume that the difference between the severity levels is also represented by a factor of 10. In summary, this difference appears between any step up in the risk matrix, either in the horizontal or in the vertical direction, i.e. when jumping from one diagonal to the neighboring one. In general, the overall risk is dominated by the risks appearing at the highest diagonal, according to the exponential scaling. The next levels constitute combined risks which are decreased by a factor of 10, 100, 1000, etc. Thus, these values represent average differences. There may be cases where neighboring risks are closer because one or both of them lie at the border to the next class.

An additional requirement in the risk management standards (8, 9) is the discrimination between hazardous situations, hazards, and harms. Harms are actual damages to humans, goods or the environment. Hazards are situations where harms may eventually occur. Hazardous situation describes a situation where humans, goods or the environment are exposed to a hazard. (8) Thus, the pure occurrence of a FP or FN case is not really a risk but a hazardous situation, since an FP or FN does not create a harm directly. For example, an FN in an ML-based test for cancer screening indicates that a harm may result. But, it does not indicate that some actual level of harm actually has occurred. This may depend on the individual development of the potential disease, i.e. whether a critical or a lower stage of disease is obtained. Thus, two different factors p_1 and p_2 constitute the probability of harm, where p_1 represents the probability of the hazard, e.g. a FP or FN case, and p_2 is the probability that a harm occurs when the hazard is given. The overall probability of harm then is $p_1 \cdot p_2$. (8) Since our approach focuses on the particular probabilities for FP and FN, e.g. $P(FN) = P \cdot FNR(s)$, i.e. the hazards, this refers to the probability p_1. Thus, the probability p_2 has to be integrated into the weight factors w_{FP} and w_{FN}, when considering the expected risk $ER(s) = w_{FP} \cdot N \cdot FPR(s) + w_{FN} \cdot P \cdot FNR(s)$. Additionally, there may be other measures, e.g. other tests or effective therapies also in later stages, which could have the potential to mitigate the risk in terms of probability or severity. These would also have to be integrated into the weights w_{FP} and w_{FN}. Even though such options were not elaborated in this paper, they can basically be addressed appropriately. Basically, such options are also feasible in the framework of normative decision theoretic framework (22).

Finally, we checked how the basic regulatory requirements apply to the use scenarios provided in section 2.3. These scenarios include substantial differences in the risk profiles. The according analysis can be found in Tab. 6. Mind that in all these use scenarios, the probabilities for the different types or errors / risks were assumed to be equal. Only the costs for the risks and subsequently the overall risk ratios differed. Additionally, a default risk ratio of $c_{FN} = 1$ was assumed for the reference scenario considered as a case of moderate risk. Within this analysis, the deviations of the risk ratio according to the reported risk aspects were roughly estimated.

Tab. 6. Analysis of use scenarios as introduced in section 2.3: impact of particular settings / risk factors on the overall risk ratio. A default risk ratio of $c_{FN} = 1$ was assumed as a reference for moderate risk levels. The deviations to this default value due to the details in the particular case were rated.

Use scenario	**implication on costs / overall risk ratio**
A. ***diagnostic test:*** ML-based system which is integrated into a screening test for a specific disease (e.g. a specific type of cancer). The actual prevalence of the disease as well as the probabilities of different types of errors / risks, i.e. $\boldsymbol{TP}$, $\boldsymbol{FN}$, $\boldsymbol{TN}$, and $\boldsymbol{FP}$, is assumed to be fixed in the following subcases.	

1. situation with very high risk in case of false negatives ($\boldsymbol{FN}$), when an early detection of the disease is missed, e.g. because it quickly develops into a critical state where the success rate of potential treatments is very limited	substantially higher costs for FN $\rightarrow c_{FN} \gg 1$
2. situation still with high risk in case of false negatives ($\boldsymbol{FN}$), because the impact of the disease basically is serious, but with an option to better detect the disease by additional tests	more moderate costs for FN, if the test is integrated as an additional measure; impact depends on the quality of the additional test
3. situation with reduced risk in case of false negatives ($\boldsymbol{FN}$), because the disease develops rather slowly and has less severe impact	moderate to low costs for FN $\rightarrow c_{FN} < 1$
4. situation with reduced risk in case of false negatives ($\boldsymbol{FN}$), like in scenario AA.3, but additionally with high risk in the case of false negatives ($\boldsymbol{FP}$), e.g. when a biopsy or another treatment needs to be performed in the case of positively predicted case (i.e. $\boldsymbol{TP}$ and $\boldsymbol{FP}$), which may cause substantial harm to the patient	substantially higher costs for FP $\rightarrow c_{FN} \ll 1$ (if not counter-balanced by other types of harm)
B. ***quality inspection:*** ML-based quality assurance system for identifying deficiencies in surgical instruments before they get delivered. It is assumed that the same ratio relationships between positive (instrument has a defect) and negative cases (instrument has no defect) as well as error cases (i.e. $\boldsymbol{TP}$, $\boldsymbol{FN}$, $\boldsymbol{TN}$, and $\boldsymbol{FP}$) is given as in use scenario A.	
1. situation where instruments with a missed detection of a defect ($\boldsymbol{FN}$) will be delivered directly to a hospital and may cause serious harm to a patient when applied in the treatment procedure	potentially high costs for FN, if defect cannot be detected otherwise $\rightarrow c_{FN} > 1$
2. situation as in case B1, but this time including an additional check in the hospital which substantially lowers the probability and/or severity of the potential harm of $\boldsymbol{FN}$ cases	Substantially lower costs for FN in comparison to scenario B1 $\rightarrow c_{FN} < 1$
3. situation where the quality assurance step is designed to identify defects in an early production step and eliminate the particular instrument to reduce further financial costs, caused by $\boldsymbol{FP}$. In this case, it is considered that additional quality steps are included to keep the $\boldsymbol{FN}$ rate at an appropriate level, e.g. additional visual inspections or tests, which reduce the risk of delivering defect	only limited impact on clinical aspects, but the company should be interested to do a cost-based assessment due to financial reasons

instruments / producing harm on the patient to a low and acceptable level.	

As a result, it can be recognized that there are several situations which lead to risk ratios c_{FN} which may considerably deviate from $c_{FN} = 1$. This includes deviations in either direction, e.g. increases of c_{FN} due to higher risks for FN cases as well as decreases of c_{FN} due to lower risks for FN cases as well as higher risks for FN cases. Mind that one step up in the risk matrix usually corresponds with an increase of the risk ratio by a factor of 10. Additionally, there are cases where the impact depends on other measures (e.g. additional tests or the impact of specific treatment options). In these cases, the chain of effects needs to be considered in order to obtain a proper estimation of the overall risk ratio. This would lead to a decision making process with a deeper structure of dependencies, which is not directly addressed in this paper.

One critical aspect in this process is the question how to get to appropriate probabilities and costs for the particular risks. If they are known, they should be integrated into the evaluation of the ML-based models according to the discussed requirements in the MDR (6) and risk management standard (8). If they are not known, the question is whether and to what detail they actively need to be determined during the development phase. This may depend on the particular use case and thus, needs to be analyzed on this level. As an alternative, it may be possible or required to collect data during the operation period of the device, within the post market surveillance activities. Thus, an incremental strategy for the more detailed determination of risk factors may be feasible. In general, risk management should be considered and implemented as a continuing process. According to the MDR (6) as well as the proposed AI Act (7), it is also necessary to thoroughly follow up the results of the operation phase and eventually update the device, if the risk profile substantially changes. As already mentioned, it is allowed to perform a semi-quantitative or even qualitative assessment of the risks, according to (8, 9). This allows that certain levels of risk can be grouped together and categorized with respect to the probability as well as the severity level. This renders the assessment of risks more practicable.

4 Discussion

Within this paper, we demonstrated the necessity as well as the impact of a risk-based approach for the evaluation of ML-based medical devices, in particular for classification tasks.

4.1 Topic A – Utilization of risk-based performance metrics in recent publications

With respect to topic A, we showed that risk-based approaches currently do not play a substantial role in the scientific literature, when assessing the performance of ML-based classification models. Basically, standard metrics like BA, $F1$ score, or MCC are applied for this, according to the performed non-exhaustive literature research for an exemplary time period. Risk-based aspects are only integrated / reported in a low percentage of papers. When we counted the publications, which addressed risk prediction as the main application, as positive results, we got 3 out of 30 cases, i.e. 10%, with a 95% CI of 0.24, in the best case. When we excluded these cases fully, we got down to 0 out of 27 cases, with a 95% CI of 0.11. In any case, the application of risk-based approaches was very limited and restricted to cases where risk prediction was a main topic itself.

4.2 Topic B – Impact of risk factors into performance metrics

With respect to topic B, an approach for integrating risk factors into the evaluation of ML-based classification models was provided. In particular, dedicated weights were integrated for the different types of errors (false positives – FP and false negatives – FN) into the balanced accuracy (BA) metric as a standard performance measure. This resulted in an evaluation of ML classification models in terms of the expected risk ER respectively $\widetilde{ER}$. It was demonstrated that ER is equivalent to a performance metric, which is a weighted version of BA. Thus, this metric was subsequently called Weighted Balanced Accuracy (WBA). An artificial error distribution based on modified Gaussian distributions was utilized to analyze the impact of different risk ratios on the resulting overall expected risk. It was demonstrated, that the relative increase with respect to $\widetilde{ER}$ for the analyzed parameter settings increases up to 198% for risk ratios c_{FN} of 0.1 and 10.0, i.e. when the weights for the different types of errors FP and FN differ by such a factor. This relative increase refers to the situation, when an unweighted threshold selection (i.e. risk ratio $c_{FN} = 1$) would have been performed instead of the actual risk ratio. Risk ratios c_{FN} of 0.1 and 10.0 represent important benchmarks since they typically corresponds with an de-/increase of one level in the risk matrix, as it is often applied for medical devices according to (9). For risk ratios in the range between 0.5 and 2.0, the increase in $\widetilde{ER}$ remains lower than 12%, in our example.

4.3 Topic C – Integration into the development process for ML-based medical devices

With respect to topic C, the impact of these findings was analyzed in relationship to the regulatory requirements for the development of AI-based medical devices as given by the corresponding regulations and standards. In particular, this referred to the situation in the EU, with the MDR (6) as the main regulation for medical devices and the ISO 14971 (8) as the relevant standard for risk management. This was accompanied by the technical report ISO/TR 24791 (9) as a guidance for applying (8) as well as the proposed AI Act of the EU (7), which probably has to be applied for many AI-based medical devices in the future, in its then final version. It was demonstrated, that a neutral risk profile (with overall risk ratio $= 1$) basically requires, that the probability and severity of a risk have a reciprocal relationship, i.e. their product equals 1 when using a multiplicative approach for combining severity and probability levels. Since the latter are often given in exponential steps, the severity levels would need to have the same increase to achieve a balanced situation. Using exemplary application scenarios, we demonstrated that deviations from a reference scenario (considered as a neutral case) can occur in either direction. Since an increase of the risk ratio by a factor 10 typically refers to an increase of one level in the risk matrix, the range of risk ratios used in this paper are considered to represent reasonable scenarios for such applications. Thus, a risk-based evaluation of AI-based medical devices is required by the regulations and standards and needs to be considered in the definition of appropriate, use-case specific performance metrics.

4.4 Relation to existing approaches

In the literature, there already are some approaches to include costs and benefits into the evaluation of ML-based classification tasks as discussed in the introduction, see e.g. (12, 13, 15–22). Some of them apply to AI in general, some of them focus in medical applications. The approach presented in this paper utilizes basic aspects of this methodology, in particular within the framework of normative decision theory, and applies it to the risk-based development of medical devices. It substantially extends the preliminary results provided in (23).

Before we summarize the major findings of this paper, we do a delimination. Our paper does not address all levels of integration. For example, it does not include the costs for the correctly assigned cases. Additionally, it does not present cases where the decision has to follow a deeper structure of decisions, e.g. regarding the different probabilities and severities of developing a serious disease in the case of missed diagnosis, i.e. FN cases, or the integration of risk mitigation measure, like performing additional tests to safeguard a diagnosis or other measures to reduce the impact of a missed diagnosis. In decision / utility theory, such deeper structures can e.g. be addressed using influence diagrams (16). Additionally, our paper does not take different, non-linear ratings into account which e.g. represent a stronger risk averse behavior, i.e. over proportionately avoid risks. In particular, such extensions can be applied to deal with situations where combined risk values are not calculated by a multiplicative approach but another type of combination. Further on, more sophisticated methods regarding the impact of uncertainties, e.g. in terms of uncertainty aversion, as well as their treatment, e.g. using the value of information approach, were not addressed. (16). This could e.g. be used to include the detectability of specific errors and risks in the calculation as well as the potential costs to obtain further valuable information, e.g. about a certain disease or therapy using additional diagnostic tests.

Even though such factors are not included in this paper, our basic approach can be extended into this direction in future steps as it is compatible with the methodology of decision theory. However, the proposed methodology provides basic ingredients for the integration of risk factors into the evaluation of ML-based classification models. Based on this, important regulatory requirements can be addressed as given in (8).

The utilization of application-specific risk factors also has some challenges. First of all, the reliable assessment of probabilities and the definition of appropriate costs / weights for the different risks can be problematic. In particular, it often has to be defined how serious / critical harms should be balanced with other types of impact, e.g. additional personal burdens or costs. For balancing critical harms or even deaths with costs, the quality-adjusted life years (QALY) approach can be utilized. It basically relates to the question how much money persons are willing to spend to reach or maintain a certain level of health. (21, 59) These costs have to be coupled with the probabilities, which are also often unknown during development. Another option is the usage of micromorts. It is based on the question how much a person is willing to accept for a lottery representing a death probability of 1 in a million. (22, 60)

To integrate risk factors into the development of products, the standard for risk management for medical devices ISO 14971 (8) allows some pragmatic simplifications. On the one hand, the probabilities may be clustered in a semi-quantitative or even qualitative way based on estimations by experts. On the other hand, the risk assessment can / should be updated after its placement on the market according to systematically acquired data from the operation phase. When both factors, i.e. probabilities and costs / severity, are available, the product of these two factors provides the combined risk ratio. This reciprocal relationship was graphically shown in Fig. 5. In terms of decision theory, the different levels of risk ratio represent a so-called preference relationship (see (16) for basic definition of preference relations). Such relationships are crucial to define situations when different parameters, i.e. different aspects of utility or costs, are balanced out. In our case, this constitutes in which situations the particular risks, e.g. risks caused by FP vs. FN cases, are balanced out. They are constituted by the iso-level lines of the preference relationship. Again, this builds a bridge between our approach and the methodology developed in decision theory.

Using application-specific performance metrics has some other limitations. The comparability of different scientific approaches or models gets more challenging. Standardized metrics have the advantage that the models can be rated according to a generally established method as emphasized e.g. in (11). Additionally, standardized metrics are examined in more detail and thus, may reflect a higher level of interpretability, in some sense. This may be increased when risk-based assessment methods include multiple factors and get more complex. But, standard metrics may also achieve a lower interpretability, in some sense. Values like specificity, sensitivity, $F1$ score, MCC are abstract numbers which are hard to understand for many people. A risk-based approach better describes the results in terms of clinical, application-specific outcomes. This provides better access to the actual use of a model, including its risks / costs as well as its benefits.

4.5 Limitations of the study

The study / methods used in this paper have some limitations. First, the analysis of scientific literature was only performed for an exemplary period of time. It does not reflect the entire state-of-the-art which risk-based approaches already were developed and how often they were applied. Second, we only used an artificial model for our analysis and not results from a model which comes from a real-world scenario with an actually trained model. This includes, that our model is continuous and also differentiable, which makes it easier to align the tangents of the ROC curve with the iso-contours of the metric. We also focused on symmetrical models for most of the analysis steps. Thus, it makes sense to apply our approach in real-world scenarios. Third, the current approach was focused on relatively simple decision cases. Only costs / risk factors for error cases and for simple types of errors were included. Additionally, these errors basically represent hazardous situations and not really risks as proposed in (8). An FN case does only represent a missed diagnosis. It indicates a potential thread but does not automatically constitute an actual harm. This would have to be addressed in deeper levels of the probabilistic decision structure.

5 Conclusion

The aim of this paper was not to provide a full-scale methodology for implementing all types of decisions. It was considered as a starting point to better address a more application-specific and value-based approach, which includes actual clinical factors like associated risks into the evaluation of ML-based medical devices. Thus, it wants to create awareness towards a more risk-based way of measuring performance, with a focus on ML-based classification tasks. Based on the results of this paper, it can be recognized that a systematic integration of risk factors into the evaluation of AI-based medical devices is necessary – from a regulatory perspective as well as for an application-specific optimization of clinical outcomes. The paper demonstrates that risk factors are currently only considered in a low percentage of scientific publications. Instead, this paper provides a basic methodology to systematically integrate risk factors into the evaluation of ML-based classification models – in compliance with current and upcoming regulatory requirements for their use in medical devices.

Abbrevations

AI - artificial intelligence

ML - machine learning

MDR - medical device regulation

EU - European Union

TP - true positives

FP - false positives

TN - true negatives

FN - false negatives

FPR - false positive rate

FPR - false negative rate

Acc - accuracy

BA - balanced accuracy

Prec - precision

Sen - sensitivity

Spec - specifity

NPV - negative predictive value

PPV - positive predictive value

F1 - F1 score

MCC - Matthews correlation coefficient

ROC - receiver operating characteristics

AUROC - area under the ROC Curve

PRC - precision-recall curve

WBA - weighted balanced accuracy

ER - expected risk

Supplementary Information

The article contains the table S1 with the documentation of excluded articles for topic A as a supplementary file.

Declarations

Ethics approval and consent to participate

Not applicable. No humans were involved.

Consent for publication

Not applicable. No personal data was included.

Availability of data and materials

Not applicable. Only artificial models and no actual data sets were used.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by the project "KISS – Künstliche Intelligenz, Services und Systeme", funded by the Federal Ministry for Education and Research and the Ministry of Science, Research and Arts Baden-Württemberg under grant nr. 16DHBKI061.

Authors's Contributions

MH: Conceptualization, Formal Analysis, Methodology, Software, Visualization, Validation, Project administration, Supervision, Writing - Original Draft

CR: Conceptualization, Conceptualization, Methodology, Project administration, Supervision, Writing - Review & Editing

Literature Cited

1. Raz M, Nguyen TC, Loh E, editors. Artificial Intelligence in Medicine: Applications, Limitations and Future Directions. 1st ed. 2022. Singapore: Springer Nature Singapore; Imprint Springer; 2022. (Springer eBook Collection).

2. Liu P-R, Lu L, Zhang J-Y, Huo T-T, Liu S-X, Ye Z-W. Application of Artificial Intelligence in Medicine: An Overview. Curr Med Sci 2021; 41(6):1105–15.

3. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. Artificial Intelligence in Healthcare 2020:25–60.

4. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542(7639):115–8.

5. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. The Lancet Digital Health 2021; 3(3):e195-e203.

6. Regulation (EU) 2017/745 of the European Parliament and and of the Council on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC: MDR; 2017.

7. Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain legislative acts: AI Act; 2021.

8. ISO. ISO 14971:2019-12 Medical devices - Application of risk management to medical devices: International Organization for Standardization; 2019 2019.

9. ISO. ISO/TR 24971:2020-06 Medical devices - Guidance on the application of ISO 14971 (ISO/TR 24971:2020): International Organization for Standardization; 2020 2020.

10. Tharwat A. Classification assessment methods. ACI 2021; 17(1):168–92.

11. Maier-Hein L, Reinke A, Godau P, Tizabi MD, Christodoulou E, Glocker B et al. Metrics reloaded: Pitfalls and recommendations for image analysis validation; 2022.

12. Vickers AJ, van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016; 352:i6.

13. van Leeuwen DA, Brümmer N. An Introduction to Application-Independent Evaluation of Speaker Recognition Systems. In: Müller C, editor. Speaker classification: Fundamentals, features, and methods. Berlin: Springer; 2007. p. 330–53 (Lecture Notes in Computer Science; vol. 4343).

14. Whang SE, Lee J-G. Data collection and quality challenges for deep learning. Proc. VLDB Endow. 2020; 13(12):3429–32.

15. Schwendicke F, Rossi JG, Göstemeyer G, Elhennawy K, Cantu AG, Gaudin R et al. Cost-effectiveness of Artificial Intelligence for Proximal Caries Detection. J Dent Res 2021; 100(4):369–76.

16. Straub D, Welpe I. Decision-Making Under Risk: A Normative and Behavioral Perspective. In: Klüppelberg C, Straub D, Welpe IM, editors. Risk - a multidisciplinary introduction. Cham, Heidelberg: Springer; 2014. p. 63–93.

17. Paté-Cornell ME, Dillon RL. The Respective Roles of Risk and Decision Analyses in Decision Support. Decision Analysis 2006; 3(4):220–32.

18. Borgonovo E, Cappelli V, Maccheroni F, Marinacci M. Risk analysis and decision theory: A bridge. European Journal of Operational Research 2018; 264(1):280–93.

19. Baker SG, Cook NR, Vickers A, Kramer BS. Using relative utility curves to evaluate risk prediction. J R Stat Soc Ser A Stat Soc 2009; 172(4):729–48.

20. Rousson V, Zumbrunn T. Decision curve analysis revisited: overall net benefit, relationships to ROC curve analysis, and application to case-control studies. BMC Med Inform Decis Mak 2011; 11:45.

21. Felder S, Mayrhofer T. Medical decision making: A health economic primer. Second edition. Berlin: Springer; 2017.

22. Russell SJ, Norvig P. Artificial intelligence: A modern approach. Fourth edition, global edition. Harlow: Pearson; 2022. (Pearson Series in Artificial Intelligence).

23. Haimerl M. Risk-based Assessment of ML-based Medical Devices. In: Upper Rhine Artificial Intelligence (URAI) Conference: Conference Proceedings. Furtwangen University; 2022. p. 146–50.

24. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968; 70(4):213–20.

25. US Food and Drug Administration, editor. Guidance for the Content of Premarket Submissions for Software Contained in Medical Devices.

26. Neumann J von. Theory of games and economic behavior. 60. anniversary ed., 4. print., and 1. paperb. print. Princeton, NJ: Princeton University Press; 2007. (Princeton Classic Editions Ser). Available from: URL: https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=1092486.

27. Ozcan I, Aydin H, Cetinkaya A. Comparison of Classification Success Rates of Different Machine Learning Algorithms in the Diagnosis of Breast Cancer. Asian Pac J Cancer Prev 2022; 23(10):3287–97.

28. Garavand A, Salehnasab C, Behmanesh A, Aslani N, Zadeh AH, Ghaderzadeh M. Efficient Model for Coronary Artery Disease Diagnosis: A Comparative Study of Several Machine Learning Algorithms. J Healthc Eng 2022; 2022:5359540.

29. ElSeddawy AI, Karim FK, Hussein AM, Khafaga DS. Predictive Analysis of Diabetes-Risk with Class Imbalance. Comput Intell Neurosci 2022; 2022:3078025.

30. Kasim S, Malek S, Cheen S, Safiruz MS, Ahmad WAW, Ibrahim KS et al. In-hospital risk stratification algorithm of Asian elderly patients. Sci Rep 2022; 12(1):17592.

31. Aldhyani THH, Alsubari SN, Alshebami AS, Alkahtani H, Ahmed ZAT. Detecting and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models. Int J Environ Res Public Health 2022; 19(19).

32. Wu J, Li Y, Yin L, He Y, Wu T, Ruan C et al. Automated assessment of balance: A neural network approach based on large-scale balance function data. Front Public Health 2022; 10:882811.

33. Preto AJ, Matos-Filipe P, Mourão J, Moreira IS. SYNPRED: prediction of drug combination effects in cancer using different synergy metrics and ensemble learning. Gigascience 2022; 11.

34. González-Cebrián A, Borràs-Ferrís J, Ordovás-Baines JP, Hermenegildo-Caudevilla M, Climente-Marti M, Tarazona S et al. Machine-learning-derived predictive score for early estimation of COVID-19 mortality risk in hospitalized patients. PLoS One 2022; 17(9):e0274171.

35. He J, Li J, Jiang S, Cheng W, Jiang J, Xu Y et al. Application of machine learning algorithms in predicting HIV infection among men who have sex with men: Model development and validation. Front Public Health 2022; 10:967681.

36. Milara E, Gómez-Grande A, Tomás-Soler S, Seiffert AP, Alonso R, Gómez EJ et al. Bone marrow segmentation and radiomics analysis of 18FFDG PET/CT images for measurable residual disease assessment in multiple myeloma. Comput Methods Programs Biomed 2022; 225:107083.

37. Emakhu J, Monplaisir L, Aguwa C, Arslanturk S, Masoud S, Nassereddine H et al. Acute coronary syndrome prediction in emergency care: A machine learning approach. Comput Methods Programs Biomed 2022; 225:107080.

38. Haq EU, Jianjun H, Huarong X, Li K, Weng L. A Hybrid Approach Based on Deep CNN and Machine Learning Classifiers for the Tumor Segmentation and Classification in Brain MRI. Comput Math Methods Med 2022; 2022:6446680.

39. Movahed RA, Rezaeian M. Automatic Diagnosis of Mild Cognitive Impairment Based on Spectral, Functional Connectivity, and Nonlinear EEG-Based Features. Comput Math Methods Med 2022; 2022:2014001.

40. Templeton JM, Poellabauer C, Schneider S. Classification of Parkinson's disease and its stages using machine learning. Sci Rep 2022; 12(1):14036.

41. Zou Y, Shi Y, Sun F, Liu J, Guo Y, Zhang H et al. Extreme gradient boosting model to assess risk of central cervical lymph node metastasis in patients with papillary thyroid carcinoma: Individual prediction using SHapley Additive exPlanations. Comput Methods Programs Biomed 2022; 225:107038.

42. van Tran, Saad T, Tesfaye M, Walelign S, Wordofa M, Abera D et al. Helicobacter pylori (H. pylori) risk factor analysis and prevalence prediction: a machine learning-based approach. BMC Infect Dis 2022; 22(1):655.

43. Maskew M, Sharpey-Schafer K, Voux L de, Crompton T, Bor J, Rennick M et al. Applying machine learning and predictive modeling to retention and viral suppression in South African HIV treatment cohorts. Sci Rep 2022; 12(1):12715.

44. Mabrouk A, Dahou A, Elaziz MA, Díaz Redondo RP, Kayed M. Medical Image Classification Using Transfer Learning and Chaos Game Optimization on the Internet of Medical Things. Comput Intell Neurosci 2022; 2022:9112634.

45. Khan W, Zaki N, Masud MM, Ahmad A, Ali L, Ali N et al. Infant birth weight estimation and low birth weight classification in United Arab Emirates using machine learning algorithms. Sci Rep 2022; 12(1):12110.

46. Ho TKK, Gwak J. Feature-level ensemble approach for COVID-19 detection using chest X-ray images. PLoS One 2022; 17(7):e0268430.

47. Eissa NS, Khairuddin U, Yusof R. A hybrid metaheuristic-deep learning technique for the pan-classification of cancer based on DNA methylation. BMC Bioinformatics 2022; 23(1):273.

48. Salimpour S, Kalbkhani H, Seyyedi S, Solouk V. Stockwell transform and semi-supervised feature selection from deep features for classification of BCI signals. Sci Rep 2022; 12(1):11773.

49. Berenguer-Vidal R, Verdú-Monedero R, Morales-Sánchez J, Sellés-Navarro I, Kovalyk O, Sancho-Gómez J-L. Decision Trees for Glaucoma Screening Based on the Asymmetry of the Retinal Nerve Fiber Layer in Optical Coherence Tomography. Sensors (Basel) 2022; 22(13).

50. Dritsas E, Trigka M. Stroke Risk Prediction with Machine Learning Techniques. Sensors (Basel) 2022; 22(13).

51. Ahmad S, Ullah T, Ahmad I, Al-Sharabi A, Ullah K, Khan RA et al. A Novel Hybrid Deep Learning Model for Metastatic Cancer Detection. Comput Intell Neurosci 2022; 2022:8141530.

52. Goñi M, Basu N, Murray AD, Waiter GD. Brain predictors of fatigue in rheumatoid arthritis: A machine learning study. PLoS One 2022; 17(6):e0269952.

53. Dubol M, Stiernman L, Wikström J, Lanzenberger R, Neill Epperson C, Sundström-Poromaa I et al. Differential grey matter structure in women with premenstrual dysphoric disorder: evidence from brain morphometry and data-driven classification. Transl Psychiatry 2022; 12(1):250.

54. Hidayat SN, Julian T, Dharmawan AB, Puspita M, Chandra L, Rohman A et al. Hybrid learning method based on feature clustering and scoring for enhanced COVID-19 breath analysis by an electronic nose. Artif Intell Med 2022; 129:102323.

55. Baskozos G, Themistocleous AC, Hebert HL, Pascal MMV, John J, Callaghan BC et al. Classification of painful or painless diabetic peripheral neuropathy and identification of the most powerful predictors using machine learning models in large cross-sectional cohorts. BMC Med Inform Decis Mak 2022; 22(1):144.

56. Shakhovska N, Yakovyna V, Chopyak V. A new hybrid ensemble machine-learning model for severity risk assessment and post-COVID prediction system. Math Biosci Eng 2022; 19(6):6102–23.

57. International Organization for Standardization. DIN EN ISO 13485:2016 Medical devices - Quality management systems - Requirements for regulatory purposes (ISO_13485:2016); Deutsche Fassung EN_ISO_13485:2016_+ AC:2018_+ A11:2021: International Organization for Standardization.

58. Kirkire MS, Rane SB, Jadhav JR. Risk management in medical product development process using traditional FMEA and fuzzy linguistic approach: a case study. J Ind Eng Int 2015; 11(4):595–611.

59. Weinstein MC, Torrance G, McGuire A. QALYs: the basics. Value Health 2009; 12 Suppl 1:S5-9.

60. Howard RA. Microrisks for medical decision analysis. Int J Technol Assess Health Care 1989; 5(3):357–70.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- RiskbasedEvaluationofMLbasedClassificationSupplementS1.docx