Using machine learning to improve the diagnostic accuracy of the modified Duke/ESC 2015 criteria in patients with suspected prosthetic valve endocarditis

ten Hove, D.; Slart, R. H.J.A.; Glaudemans, A. W.J.M.; Postma, D. F.; Gomes, A.; Swart, L. E.; Tanis, W.; van Geel, P. P.; Mecozzi, G.; Budde, R. P.J.

Published in:
European Journal of Nuclear Medicine and Molecular Imaging

DOI:
10.1007/s00259-024-06774-y

European Journal of Nuclear Medicine and Molecular Imaging (2024) 51:3924–3933
https://doi.org/10.1007/s00259-024-06774-y

ORIGINAL ARTICLE

Using machine learning to improve the diagnostic accuracy of the modified Duke/ESC 2015 criteria in patients with suspected prosthetic valve endocarditis – a proof of concept study

D. ten Hove[1,2] · R. H. J. A. Slart[1,3] · A. W. J. M. Glaudemans[1] · D. F. Postma[4] · A. Gomes[2] · L. E. Swart[5] · W. Tanis[6] · P. P. van Geel[7] · G. Mecozzi[8] · R. P. J. Budde[9] · K. Mouridsen[1,10] · B. Sinha[2]

Received: 14 November 2023 / Accepted: 17 May 2024 / Published online: 21 June 2024

Abstract

Introduction Prosthetic valve endocarditis (PVE) is a serious complication of prosthetic valve implantation, with an estimated yearly incidence of at least 0.4-1.0%. The Duke criteria and subsequent modifications have been developed as a diagnostic framework for infective endocarditis (IE) in clinical studies. However, their sensitivity and specificity are limited, especially for PVE. Furthermore, their most recent versions (ESC2015 and ESC2023) include advanced imaging modalities, e.g., cardiac CTA and [^{18}F]FDG PET/CT as major criteria. However, despite these significant changes, the weighing system using major and minor criteria has remained unchanged. This may have introduced bias to the diagnostic set of criteria. Here, we aimed to evaluate and improve the predictive value of the modified Duke/ESC 2015 (MDE2015) criteria by using machine learning algorithms.

Methods In this proof-of-concept study, we used data of a well-defined retrospective multicentre cohort of 160 patients evaluated for suspected PVE. Four machine learning algorithms were compared to the prediction of the diagnosis according to the MDE2015 criteria: Lasso logistic regression, decision tree with gradient boosting (XGBoost), decision tree without gradient boosting, and a model combining predictions of these (ensemble learning). All models used the same features that also constitute the MDE2015 criteria. The final diagnosis of PVE, based on endocarditis team consensus using all available clinical information, including surgical findings whenever performed, and with at least 1 year follow up, was used as the composite gold standard.

Results The diagnostic performance of the MDE2015 criteria varied depending on how the category of 'possible' PVE cases were handled. Considering these cases as positive for PVE, sensitivity and specificity were 0.96 and 0.60, respectively. Whereas treating these cases as negative, sensitivity and specificity were 0.74 and 0.98, respectively. Combining the approaches of considering possible endocarditis as positive and as negative for ROC-analysis resulted in an excellent AUC of 0.917. For the machine learning models, the sensitivity and specificity were as follows: logistic regression, 0.92 and 0.85; XGBoost, 0.90 and 0.85; decision trees, 0.88 and 0.86; and ensemble learning, 0.91 and 0.85, respectively. The resulting AUCs were, in the same order: 0.938, 0.937, 0.930, and 0.941, respectively.

Discussion In this proof-of-concept study, machine learning algorithms achieved improved diagnostic performance compared to the major/minor weighing system as used in the MDE2015 criteria. Moreover, these models provide quantifiable certainty levels of the diagnosis, potentially enhancing interpretability for clinicians. Additionally, they allow for easy incorporation of new and/or refined criteria, such as the individual weight of advanced imaging modalities such as CTA or [^{18}F]FDG PET/CT. These promising preliminary findings warrant further studies for validation, ideally in a prospective cohort encompassing the full spectrum of patients with suspected IE.

Keywords Endocarditis · Machine learning · Modified Duke/ESC 2015 criteria

K. Mouridsen and B. Sinha Equally shared authorship to this work.

Extended author information available on the last page of the article

Introduction

Infective endocarditis (IE) is a serious condition causing substantial morbidity and mortality and a well known potential complication of prosthetic valve implantation. The risk of developing IE after prosthetic valve implantation is estimated to be at least 0.4-1.0% per patient-year [1, 2]. Simultaneously, disease severity increases with prosthetic valves being present, with a reported in-hospital mortality rate of 19.9% [3]. A timely and accurate diagnosis of prosthetic valve endocarditis (PVE) is vital for determining the best treatment plan and patient outcome. However, the clinical presentation of PVE is often variable and non-specific, making it challenging to diagnose.

In 1994 the Duke criteria were introduced as a framework for diagnosing endocarditis and although they were primarily intended for use in epidemiologic studies, they have also been used extensively in clinical settings. Since their inception, the Duke criteria have undergone four major modifications [4–6]. The second modification was made by the European Society of Cardiology (ESC) in 2015 [5] and in 2023 two other updates were published: one by the ESC and another by the International Society of Cardiovascular Infectious Diseases (ISCVID) [6, 7]. The 2015 ESC and 2023 ISCVID and ESC modifications of the Duke criteria, among other changes, introduced advanced imaging modalities such as cardiac Computed Tomography Angiography (Cardiac CTA) and [^{18}F]fluorodeoxyglucose positron emission tomography/computed tomography ([^{18}F]FDG PET/CT) as major criteria, while leaving the original weighing system of major and minor criteria unchanged. This may have introduced bias to the endocarditis criteria. This is due to the fact that the diagnostic system applies weight to all imaging modalities as if they were echocardiography, even though multimodality imaging has additional diagnostic power over echocardiography alone. The risk of this is that findings from these powerful diagnostic tools are not valued correctly for the diagnosis of PVE. And while the Duke criteria and their modifications clearly state that they are not meant to replace clinical judgment, it is still beneficial to strive to eliminate sources of potential bias from the endocarditis criteria as much as possible.

Additionally, the original Duke criteria and their subsequent modifications all allow for a ‘possible’ endocarditis designation. In clinical practice a considerable number of patients receive this designation for the diagnosis, which limits the clinical utility of the criteria. The overall sensitivity and specificity of the modified Duke criteria for IE is approximately 80%, and for patients with prosthetic heart valves or cardiac implanted electronic devices this is likely lower [8]. Recent studies regarding the exact sensitivity, specificity of the modified Duke/ESC (MDE2015) criteria and the rate at which they result in an unclear or ‘possible’ diagnosis are scarce, and should be interpreted with caution [9].

Machine learning algorithms are increasingly recognized as valuable diagnostic tools in the medical field, with applications spanning a wide range of diagnostic and prognostic applications [10]. Decision tree models are an early form of machine learning, but they continue to be popular due to their interpretability and ease of implementation [11]. The Lasso logistic regression model, a variant of logistic regression that utilizes L1 regularization, has demonstrated its utility in various studies [12, 13]. XGBoost is another model that has garnered attention in recent years. It is an implementation of decision trees which uses gradient boosting to increase its predictive power and it has shown promise in predicting diverse medical conditions [14–16]. In fact, in a recent study, it has been shown that diagnosis of aortic PVE can be substantially enhanced by using a machine learning model that includes automated segmentation and radiomics for interpretation of PET/CT as improved major criterion for the ESC-2015 criteria. The improved performance positively impacted on overall diagnostic accuracy [17].

The aim of this study was to improve the diagnostic accuracy of the (MDE2015) criteria through application of machine learning algorithms, while using the same diagnostic features that also comprise the MDE2015 criteria. Instead of the traditional division between major and minor criteria, our approach was to weigh the different MDE2015 criteria in a data-driven manner, using the aforementioned machine learning algorithms. To the best of our knowledge, this is the first study using machine learning in order to optimize application of the MDE2015 criteria. The goal of this was assuring that the different features comprising the MDE2015 criteria were used in accordance with their predictive value for the diagnosis of PVE. We hypothesized that this approach would make better use of the multiple diagnostic advancements that have found their way into clinical practice for diagnosing PVE over the past decades, and therefore outperform the prediction system of major and minor criteria as used by the MDE2015 endocarditis criteria, both in terms of interpretability and diagnostic accuracy.

Methods

For this study, data were used from an earlier published multicentre study [18]. In this study, prosthetic valve recipients from six cardiothoracic centres in the Netherlands were included if they had undergone [^{18}F]FDG PET/CT imaging for suspected PVE. All clinical data and the outcomes of the imaging modalities that were included in the 2015 ESC guidelines (i.e. cardiac CTA and [^{18}F]FDG PET/CT) were

available, with the exception of vascular and immunologic phenomena, which were not always documented. All available clinical data was combined to calculate the clinical MDE2015 criteria for all study subjects. Histopathologic criteria were not used to predict the disease, but when available, they were used by the endocarditis team to establish the final diagnosis.

The original paper used the modified Duke criteria from prior to the ESC 2015 update, as proposed by Li et al. [4]. The reason for this was that this study aimed to evaluate the impact of [^{18}F]FDG PET/CT on the diagnosis of PVE. Thus, incorporating findings of PET/CT into the Duke criteria carried a potential risk of bias at that stage. However, findings from PET/CT and CTA were available in the original dataset. When these imaging findings were incorporated in the MDE2015 criteria, 6 patients initially classified as negative were reclassified as "possible" PVE, and likewise 3 were reclassified from "possible" to "definite" PVE (Table 1).

The algorithms for data-driven prediction of PVE were written in Python 3.9® [Python Software Foundation, DE, USA]. To minimize the risk of overfitting the predictive models, internal cross-validation was performed using a repeated stratified K-fold to divide the data in training and test sets, using 4 folds per iteration, repeated for a total of 20 times. A Lasso logistic regression, a decision tree model and a gradient boosted decision tree model (XGBoost) were used for predicting PVE using the MDE2015 features. Additionally, these models were combined in an ensemble learning model to give one overall prediction of the disease. The maximum number of iterations to reach convergence was set to 1000. For XGboost and the decision tree classifier, the maximum depth of the trees was set to 4. For the ensemble learning classifier, the probability of PVE presence according to all the other machine learning models was averaged without first dichotomizing these predictions. Additionally, a single run of K-fold was performed to allow for a tentative statistical comparison between the models and the MDE2015 criteria. This analysis focused on the subgroup of the patient population with a possible endocarditis according to the MDE2015 criteria. For all models, a probability cut-off of 0.5 was used to classify a patient as having confirmed or rejected PVE.

The performance of the models based on their binary predictions of disease presence or absence across the repeated K-fold iterations was used for calculating their averaged performance metrics. We included accuracy, precision, F1-score, sensitivity, and specificity. All aforementioned algorithms give a probabilistic prediction for the presence of PVE. This probability-per-case prediction was also averaged over the 20 repetitions of K-fold and this was used for the Receiver Operating Curve (ROC) and Area Under the Curve (AUC) analysis curves.

The overall predictive performance of the machine learning models was compared to the prediction of the disease

Table 1 Patient demographics and clinical characteristics, by presence of PVE [18]

Demographics	PVE confirmed (*n* = *80*)	PVE rejected (*n* = *80*)	Total (*n* = *160*)	*p*-value
Age, median [IQR], y	52 [35–68]	68 [53–75]	62 [43–73]	**< 0.001**
Sex (male), n (%)	49 (61.3)	59 (73.8)	108 (67.5)	0.128
BMI, mean ± SD, kg/m^2	24.9 ± 5.4	25.6 ± 4.6	25.2 ± 5.0	0.392
Diabetes mellitus, n (%)	10 (13)	13 (16.3)	23 (14.4)	0.653
Prior history of IE, n (%)	*15* (19)	18 (23)	33 (20.6)	0.563
Multiple PV present, n (%)	9 (11.3)	12 (15)	21 (13.1)	0.640
CIED present, n (%)	8 (12.7)	9 (15.8)	17 (14.2)	0.794
(Missing data)	*(17 missing)*	*(23 missing)*	*(40 missing)*	
Surgery performed, n (%)	44 (55)	1 (1.3)	45 (28.1)	**< 0.001**
MDE2015 Criteria n (%)	74 (92.5)	12 (15)	86 (54)	**< 0.001**
Imaging (major)	55 (68.8)	18 (17.5)	73 (46)	**< 0.001**
Blood cultures (major)	80 (100)	80 (100)	160 (100)	**< 0.001**
Predisposition	58 (72.5)	27 (33.8)	85 (53)	**< 0.001**
Fever	8 (10)	0 (0)	8 (5)	**0.007**
Vascular phenomena	5 (6.3)	0 (0)	5 (3)	0.059
Immunologic phenomena	8 (10)	4 (5)	12 (7.5)	0.369
Microbiologic evidence (minor)				
MDE2015 classification, n (%)				
Rejected	3 (3.8)	48 (60)	51 (31.9)	**< 0.001**
Possible	18 (22.5)	30 (37.5)	48 (30)	
Definitive	59 (73.8)	2 (2.5)	61 (38.1)	
Mortality in follow-up, n (%)	10 (12.5)	11(13.8)	21 (13.1)	1.000

Legend % percentage of patient population, CIED: cardiac implanted electronic devices, IQR: interquartile range, n: number of patients, MDE2015: modified Duke criteria according to the 2015 European Society of Cardiology guideline [5], PV: prosthetic valve, PVE: prosthetic valve endocarditis, SD: standard deviation, y: years

according to the MDE2015 criteria as described in the 2015 ESC guidelines [5]. As previously reported, the final diagnosis of PVE in this cohort was based on multidisciplinary consensus from the endocarditis team, using all the available clinical data including surgical findings whenever performed and a minimum of 1 year of follow-up [18]. This was used as the gold standard for the diagnosis.

The MDE2015 criteria allow for a 'possible' designation for IE. This presents a challenge whenever they are compared to any predictors with a dichotomous outcome, as the MDE2015 criteria do not make a definitive prediction of the disease's presence or absence for 'possible' cases. We addressed this by calculating the MDE2015 criteria diagnostic accuracy in four scenarios: (1) assuming a 50% prediction rate for 'possible' cases, both for those with and without the disease; (2) treating all 'possible' cases as positive, (3) treating all 'possible' cases as negative; and (4) excluding all 'possible' cases. We also drew a ROC-curve and calculated the corresponding AUC for the MDE2015 criteria using scenario's 2 and 3 to give an visualization of their performance.

Next, we conducted a feature importance analysis for all machine learning models, with exception for the ensemble learning model, since ensemble learning does not allow for this type of analysis. We used feature importance analysis to get a rough estimate of the weight the models attributed to the different features that make up the MDE2015 criteria.

Statistics

Demographic data are represented as means +/- SD for continuous variables if they were normally distributed and as median and interquartile range (IQR) if they were not normally distributed. Categorical variables are shown as frequencies. Demographic differences across the outcome variable of presence of PVE by final diagnosis were tested for statistical significance using Student's t-test or Mann-Whitney U test, depending on normality or non-normality of the data. Fischer's exact test was used for comparing categorical variables. The single run of K-fold cross-validation using four folds was also performed to test for demographic differences across the folds. This analysis was added as supplemental data for illustrative purposes, showing that K-fold cross-validation results in balanced groups.

ROC curves were obtained for both the machine learning models and the MDE2015 criteria, and they were utilized for AUC analysis. In repeated K-fold cross-validation the AUCs obtained across the repeated folds cannot directly be compared with the AUC achieved by using the MDE2015 criteria due to the interdependence of the different ROCs that were obtained through this method. Therefore no further statistical analyses were performed, and the performance of the models was added using descriptive statistics only.

A two-tailed P-value of ≤ 0.05 was considered statistically significant in all analyses. Statistical analysis was performed using IBM® SPSS 26 [IBM Corp, NY, USA].

Results

As previously reported, 160 patients with suspicion of PVE, were included in the study [18]. PVE was established as the final diagnosis by the endocarditis team in 80 (50%) of these patients and rejected in the others. A full overview of patient characteristics, stratified by their assigned K-fold group in the cross-validation, is shown in Table 1.

MDE2015 criteria

In this cohort, 48 patients (30%) were classified as 'possible' endocarditis according to the MDE2015 criteria. When the predictions in this group were considered as correct in 50% of cases, both for those with and without the disease, sensitivity and specificity were 0.85 and 0.79, respectively. When all patients with 'possible' endocarditis were considered as confirmed cases, sensitivity and specificity of the criteria were 0.96 and 0.60, respectively. When they were designated as rejected endocarditis, this resulted in a sensitivity and specificity of 0.74 and 0.98, respectively. These two scenarios (all possible endocarditis considered positive/negative) were combined for ROC analysis. This yielded an AUC of 0.917, see also Fig. 1. When only patients with either a rejected or definite diagnosis of endocarditis were considered, the MDE2015 criteria achieved a sensitivity and specificity of 0.95 and 0.96, respectively.

Machine learning models

The Lasso Logistic Regression model achieved a sensitivity and specificity of 0.93 and 0.85, respectively. The XGboost model obtained a sensitivity and specificity of 0.90 and 0.85, respectively. The decision tree classifier had a sensitivity and specificity of 0.88 and 0.86, respectively, while the ensemble learning model achieved a sensitivity and specificity of 0.91 and 0.85, respectively. The AUCs of the models were 0.938 for the Lasso Logistic Regression model, 0.937 for the XGBoost model, 0.930 for the decision tree model, and 0.941 for the ensemble Learning model. Further performance metrics are shown in Table 2 and the ROC-curves for the models are shown in Figs. 2, 3, 4 and 5.

There were substantial differences between the predictive models in how much weight was attributed to the different features comprising the MDE2015 criteria. Notably,

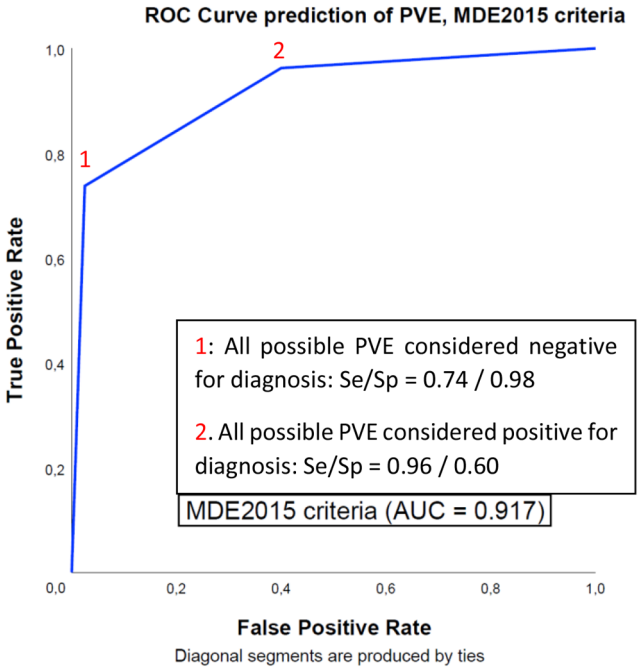

Fig. 1 ROC curves probability prediction of PVE, Modified Duke/ESC 2015 criteria. This figure provides a visualisation of the performance of the Modified Duke / ESC 2015 criteria in two different scenarios as described above. AUC: Area under the curve, ESC: European Society of Cardiology, MDE2015: Modified Duke criteria according tothe modification as proposed by the ESC in 2015, PVE: prosthetic valve endocarditis, ROC: Receiver Operating Curve, Se: Sensitivity, Sp: Specificity

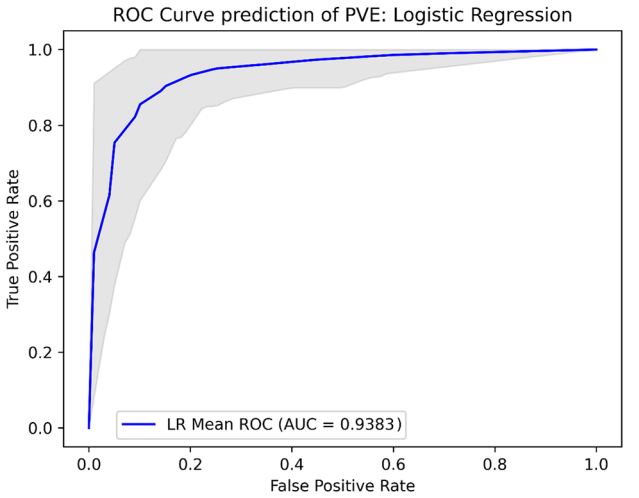

Fig. 2 ROC curves probability prediction of PVE, Logistic Regression. ROC analysis of the lasso logistic regression model, validated using repeated K-fold cross-validation. The blue line indicates the mean ROC curve across 100 iterations in repeated K-fold. The shaded area indicates the 95% confidence interval, illustrating the range of sensitivity and specificity combinations that could be encountered when applying this machine learning model across various threshold values for predicting PVE. AUC: Area under the curve, LR: logistic regression, PVE: prosthetic valve endocarditis, ROC: Receiver Operating Curve

all machine learning models attributed a large weight to (advanced) imaging modalities. With the Lasso logistic regression, the coefficient of imaging was 3.5, more than 3 times higher than that of major blood cultures. In the XGboost model, the weight of imaging was 0.91 while for the decision tree it was 0.75. An overview of feature weight per model is shown in Table 3.

In the sub-analysis of a single run K-fold, focused specifically on those patients classified as 'possible' endocarditis, the models all achieved a sensitivity and specificity of 0.94 and 0.67, respectively. The AUCs and 95% confidence intervals of the models were 0.80 (0.67–0.93) for the Lasso Logistic regression, 0.82 (0.70–0.94) for the XGBoost model, 0.80 (0.67–0.93) for the Decision Tree model and 0.81 (0.69–0.94) for the Ensemble Learning model. Tested against an AUC of 0.50, the models obtained p-values of 0.001, <0.001, 0.001 and <0.001, respectively. See also supplemental data (Fig. 1).

Discussion

This proof-of-concept study demonstrated the feasibility of machine learning models in prediction PVE with a potentially superior diagnostic accuracy compared to the MDE2015 criteria, while using the same features. Moreover, these models offer the possibility to quantify the level of uncertainty per prediction, enhancing clinical interpretability. Furthermore, they allow for easy incorporation of new and/or refined features, which would ensure that criteria based on machine learning models could easily be updated to keep them aligned with clinical practice as new and refined diagnostic tools become available. All machine

Table 2 Machine learning models performance metrics, across repeated K-fold

Metrics	Logistic regression	XGBoost	Decision trees	Ensemble learning
Accuracy, median [IQR]	0.88 [0.85-0.93]	0.88 [0.85-0.93]	0.88 [0.84-0.90]	0.88 [0.82-0.90]
Presicion, median [IQR]	0.86 [0.82-0.90]	0.86 [0.82-0.90]	0.86 [0.82-0.91]	0.86 [0.82-0.91]
F1-score, median [IQR]	0.89 [0.86-0.93]	0.88 [0.86-0.92]	0.87 [0.83-0.91]	0.89 [0.86-0.93]
Sensitivity, mean±SD	0.93±0.05	0.90±0.09	0.88±0.11	0.91±0.08
Specificity, mean±SD	0.85±0.07	0.85±0.07	0.86±0.08	0.85±0.07
AUC, mean±SD	0.94±0.03	0.94±0.04	0.93±0.04	0.94±0.03

Legend IQR: interquartile range, SD: standard deviation

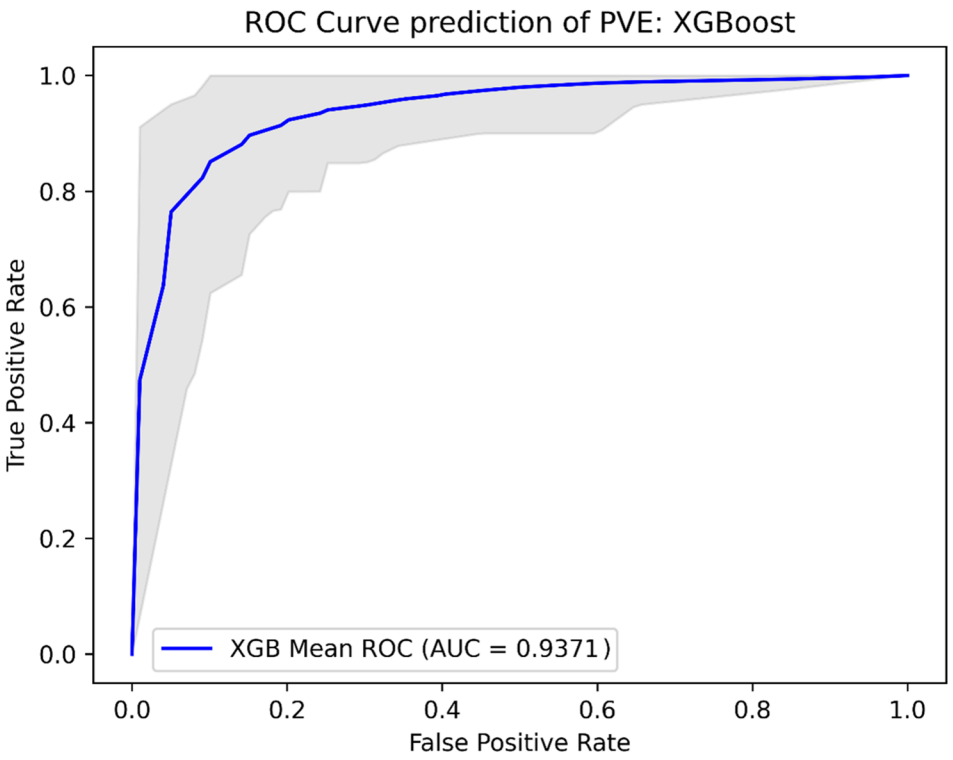

Fig. 3 ROC curves: XGBoost ROC analysis of the XGBoost model, validated using repeated K-fold cross-validation. The blue line indicates the mean ROC curve across the 100 iterations in repeated K-fold. The shaded area indicates the 95% confidence interval, illustrating the range of sensitivity and specificity combinations that could be encountered when applying this machine learning model across various threshold values for predicting PVE. AUC: Area under the curve, XGB: XGBoost, PVE: prosthetic valve endocarditis, ROC: Receiver Operating Curve

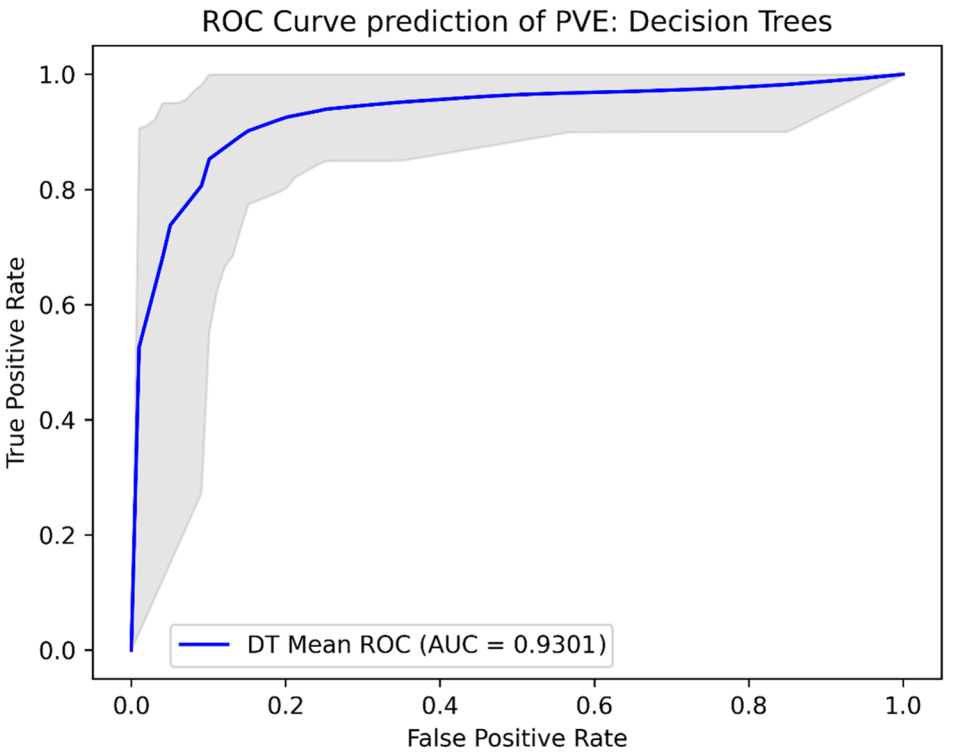

Fig. 4 ROC curves probability prediction of PVE, Decision Trees Legend ROC analysis for the decision trees model, validated using repeated K-fold cross-validation. The blue line indicates the mean ROC curve across the 100 iterations in repeated K-fold. The shaded area indicates the 95% confidence interval, illustrating the range of sensitivity and specificity combinations that could be encountered when applying this machine learning model across various threshold values for predicting PVE. AUC: Area under the curve, DT: decision trees, PVE: prosthetic valve endocarditis, ROC: Receiver Operating Curve

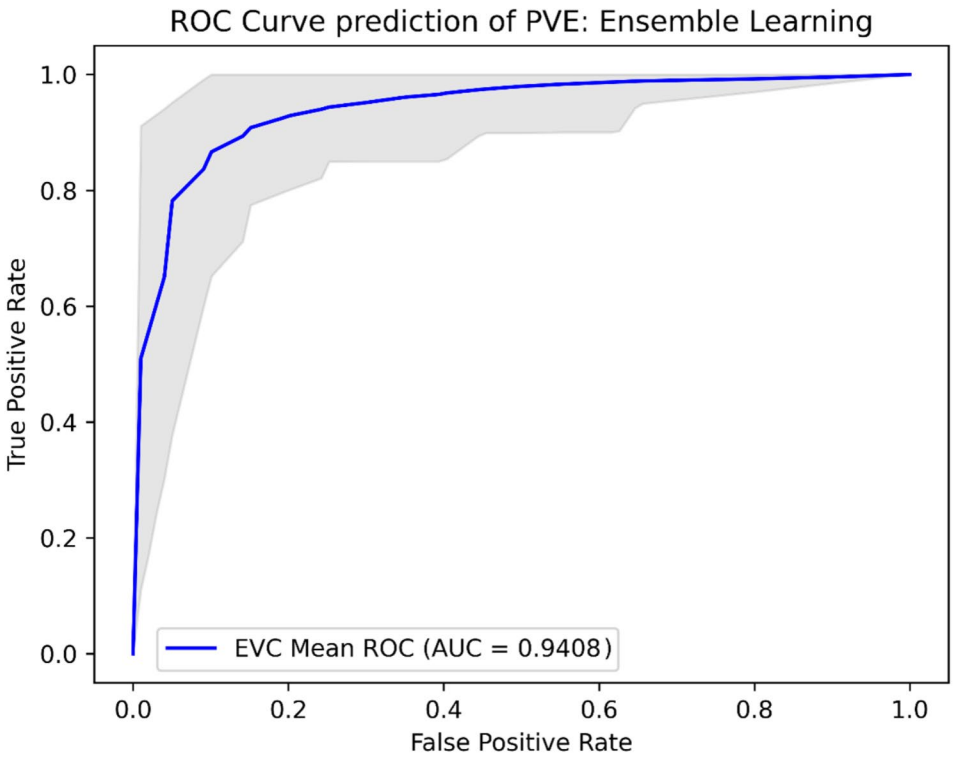

Fig. 5 ROC curves probability prediction of PVE, Ensemble Learning Legend: ROC analysis for the ensemble learning model, validated using repeated K-fold cross-validation. The blue line indicates the mean ROC curve across the 100 iterations in repeated K-fold. The shaded area indicates the 95% confidence interval, illustrating the range of sensitivity and specificity combinations that could be encountered when applying this machine learning model across various threshold values for predicting PVE. AUC: Area under the curve, EVC: ensemble learning, PVE: prosthetic valve endocarditis, ROC: Receiver Operating Curve

Table 3 Feature importance per predictive model (Repeated stratified K-fold)

Diagnostic model #	Feature importance per MDE2015 criterium[1]					
	Blood cultures (major)	Imaging	Fever	Vascular phenomena*	Immunologic Phenomena*	Microbiologic evidence (minor)
Lasso logistic regression	1.015	3.506	0.644	0.348	0.167	0.211
XGBoost	0.052	0.906	0.023	0.000	0.000	0.020
Decision trees	0.092	0.746	0.040	0.037	0.034	0.050

Legend[1]Predisposition was not included in this table as a feature, since this was present in all patients

Feature importance is measured differently in the various machine leaning models. For Logistic regression this represents the coefficients for the included features, for XGBoost it indicates total gain, while for decision trees it described the effect of a feature on the entropy in the model

* indicates features for which data was frequently missing

learning models demonstrated similar performance in terms of binary predictions. There were small differences between the models in this regard, but these are unlikely to be of clinical relevance.

Performance of the MDE2015 criteria in the current study was comparable, if not superior relative to its performance in other recent publications. In one study the MDE2015 criteria achieved an AUC of 0.87 [8] compared to 0.917 in the current study. In another recent study in which 'possible' cases were classified as negative, a sensitivity and specificity of 0.84 and 0.71 were reported [9]. In our study, the application of the MDE2015 criteria yielded a sensitivity of 0.74 and a notably higher specificity of 0.98. This illustrates the robust performance of the machine learning models, as they were benchmarked against well-established diagnostic criteria showing good performance.

Interestingly, both the XGBoost and decision tree model assigned considerable weight to the outcome of imaging while downplaying the role of blood cultures, even though the latter are also a key component in the diagnosis of PVE. In the logistic regression model, the pronounced role of imaging was also present. Here the major blood cultures contribution to the final prediction of the model was higher, but still the coefficient for imaging was more than 3 times higher than that of blood cultures. The reasons for this could be the time points of the respective features in the diagnostic process. Blood cultures are frequently the reason for suspicion of IE in patients with prosthetic valves, while imaging is applied at a later stage in the process for specific evidence of the infection intracardially. The high weight attributed to imaging could also be due to the increasing accuracy of these modalities in detecting PVE, since in the ESC guidelines from 2015 to 2023, imaging consists of findings of not

only echocardiography, but also that of cardiac CTA, [^{18}F] FDG PET/CT and leucocyte scintigraphy. Simultaneously, it might also in part reflect a high reliance of participating endocarditis teams on imaging findings for their final diagnosis, implying a level of incorporation bias.

When interpreting feature importance analysis, it is also important to stress the differences between the models. In decision tree models and XGBoost, each feature is used sequentially to make a prediction. This means that feature importance signifies the contribution of these factors to the performance of the model (e.g. total gain for XGBoost and entropy for the decision trees). In contrast, Lasso logistic regression uses all features simultaneously. In this instance, feature importance shows how heavily the different factors influence the predicted probability of disease presence in the optimized model.

Some limitations of the current study need to be addressed. For this study we used a retrospective cohort primarily investigated for the value of [^{18}F]FDG PET/CT. Given that the majority of patients were evaluated in tertiary care centres and all underwent [^{18}F]FDG PET/CT our results may not generalize to wider clinical practice. Also, our data was collected around the time of the ESC modification and therefore, could not be adapted to the recent ISCVID or 2023 ESC modifications to the Duke criteria. Furthermore, as reported in the 2018 study [18] data on vascular and immunologic phenomena were often missing in the study cohort, which means the feature weights as found in our results can illustrate how these models function in this particular cohort, but they should not be generalized to general clinical practice. Despite these limitations, the findings of this study suggest that machine learning models could enhance the utilization of available information beyond what is possible with the MDE2015 criteria.

Several considerations need to be addressed before integrating these models into clinical practice. In infective endocarditis, obtaining a definitive diagnosis is challenging if not impossible unless surgery is performed. This necessitates reliance on team consensus and follow-up. A potential concern of that is that this process could lead to bias from the participating endocarditis teams inadvertently being incorporated into the machine learning models. Moreover, key diagnostic features may be underestimated according to the models, depending on their place in the diagnostic process and their prevalence in the target population.

Medical team consensus is also often more nuanced than a binary disease presence statement. Therefore, including a level of confidence in the diagnosis could provide a more accurate benchmark for evaluating the models. Similarly, it could be beneficial to document the level of confidence associated with the different clinical signs and diagnostic tools used in patient evaluation, as these factors are not standardized: e.g., image quality may affect results of imaging modalities, antibiotic (pre-) treatment can impact results, and clinical signs may not hold the same predictive value if these are linked to a pre-existing condition. Addressing these considerations could help minimize the repeated rounding errors caused by the digitization of these diagnostic features, which would lead to more transparent and objective diagnostic decisions.

For future validation, these models should be tested on a diverse patient cohort across various healthcare settings and ideally including those with native valves, prosthetic valves and implanted cardiac devices. Hyperparameter tuning would be important for these validation studies and might show performance differences between the machine learning models used. Given a sufficiently large and diverse patient cohort to avoid overfitting, these models could also be further refined with additional predictors to provide a more nuanced understanding of the added risk of IE associated with specific types of predisposition, clinical signs or findings of the various available diagnostic modalities.

Conclusion

In this proof-of-concept study, machine learning algorithms demonstrated potentially superior performance in predicting PVE compared to the MDE2015 criteria, while using the same diagnostic features. Preliminary statistical tests were particularly promising for those with possible endocarditis according to the MDE2015 criteria. These results imply that the use of machine learning algorithms could potentially lead to improved diagnostic accuracy and interpretability for this challenging diagnosis. Future prospective validation studies are warranted to affirm whether these promising findings reflect a statistically significant improvement over the MDE criteria as they are currently used. Ideally, such studies would encompass the full spectrum of patients with suspected IE.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00259-024-06774-y.

Funding This work was supported in part by PUSH, a collaborative framework project of the University Medical Center Groningen and Siemens Healthineers. The funding source had no role in the conceptualization, analyses, writing, or publication of the article.

Data availability The data are exclusively available for the purpose of reproducing the study results upon reasonable request [18].

Declarations

Ethics approval and informed consent The study from which the data of this manuscript were derived was approved and informed consent

was waived by the local Medical Ethics Committees of all participating centers. This was in view of the retrospective nature of the study and all the procedures being performed were part of the routine care. For the current study, no new patient data were collected.

Conflict of interest DtH, AG, RS, and BS report the aforementioned institutional funding through PUSH. Additionally, BS reports grants from Beatrixoord Foundation, grants from the European Union, outside the submitted work, as well as Committee work (executive boards and working parties) on a local, national, and international level dealing with guideline development of infections. RS is an editor in this journal. RB reports institutional support to Erasmus MC from Heartflow and Siemens, outside the submitted work and speakers fees from Bayer.

References

1. Cahill TJ, Raby J, Jewell PD, Brennan PF, Banning AP, Byrne J et al. Risk of infective endocarditis after surgical and transcatheter aortic valve replacement. Heart 2022;108:639 LP – 647. https://doi.org/10.1136/heartjnl-2021-320080.
2. Hadji-Turdeghal K, Jensen AD, Bruun NE, Iversen KK, Bundgaard H, Smerup M, et al. Temporal trends in the incidence of infective endocarditis in patients with a prosthetic heart valve. Open Heart. 2023;10:e002269. https://doi.org/10.1136/openhrt-2023-002269.
3. Habib G, Erba PA, Iung B, Donal E, Cosyns B, Laroche C, et al. Clinical presentation, aetiology and outcome of infective endocarditis. Results of the ESC-EORP EURO-ENDO (European infective endocarditis) registry: a prospective cohort study. Eur Heart J. 2019;40:3222–32. https://doi.org/10.1093/eurheartj/ehz620.
4. Li JS, Sexton DJ, Mick N, Nettles R, Fowler VG, Ryan T, et al. Proposed modifications to the Duke Criteria for the diagnosis of infective endocarditis. Clin Infect Dis. 2000;30:633–8.
5. Habib G, Lancellotti P, Antunes MJ, Bongiorni MG, Casalta JP, DelZotti F et al. Guidelines for the management of infective endocarditis. 2015. https://doi.org/10.1093/eurheartj/ehv319.
6. Fowler VG, Durack DT, Selton-Suty C, Athan E, Bayer AS, Chamis AL, et al. The 2023 Duke-ISCVID Criteria for Infective endocarditis: updating the modified Duke Criteria. Clin Infect Dis. 2023. https://doi.org/10.1093/cid/ciad271.
7. Delgado V, Ajmone Marsan N, de Waha S, Bonaros N, Brida M, Burri H, et al. 2023 ESC guidelines for the management of endocarditis. Eur Heart J. 2023;44:3948–4042. https://doi.org/10.1093/eurheartj/ehad193.
8. Primus CP, Clay TA, McCue MS, Wong K, Uppal R, Ambekar S, et al. 18F-FDG PET/CT improves diagnostic certainty in native and prosthetic valve infective endocarditis over the modified Duke Criteria. J Nuclear Cardiol. 2021. https://doi.org/10.1007/s12350-021-02689-5.
9. Philip M, Tessonier L, Mancini J, Mainardi JL, Fernandez-Gerlinger MP, Lussato D, et al. Comparison between ESC and Duke Criteria for the diagnosis of prosthetic valve infective endocarditis. JACC Cardiovasc Imaging. 2020;13:2605–15. https://doi.org/10.1016/j.jcmg.2020.04.011.
10. Slart RHJA, Williams MC, Juarez-Orozco LE, Rischpler C, Dweck MR, Glaudemans AWJM, et al. Position paper of the EACVI and EANM on artificial intelligence applications in multimodality cardiovascular imaging using SPECT/CT, PET/CT, and cardiac CT. Eur J Nucl Med Mol Imaging. 2021;48:1399–413. https://doi.org/10.1007/s00259-021-05341-z.
11. Podgorelec V, Kokol P, Stiglic B, Rozman I. Decision trees: an overview of their use in medicine. J Med Syst. 2002;26:445–63. https://doi.org/10.1023/A:1016409317640.
12. Li Y, Lu F, Yin Y. Applying logistic LASSO regression for the diagnosis of atypical Crohn's disease. Sci Rep. 2022;12:11340. https://doi.org/10.1038/s41598-022-15609-5.
13. Li Y, He Y, Meng Y, Fu B, Xue S, Kang M, et al. Development and validation of a prediction model to estimate risk of acute pulmonary embolism in deep vein thrombosis patients. Sci Rep. 2022;12:649. https://doi.org/10.1038/s41598-021-04657-y.
14. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inf Decis Mak. 2019;19:211. https://doi.org/10.1186/s12911-019-0918-5.
15. Pfaff ER, Girvin AT, Bennett TD, Bhatia A, Brooks IM, Deer RR, et al. Identifying who has long COVID in the USA: a machine learning approach using N3C data. Lancet Digit Health. 2022;4:e532–41. https://doi.org/10.1016/S2589-7500(22)00048-6.
16. Chang W, Liu Y, Xiao Y, Yuan X, Xu X, Zhang S, et al. A machine-learning-based prediction method for hypertension outcomes based on Medical Data. Diagnostics. 2019;9:178. https://doi.org/10.3390/diagnostics9040178.
17. Godefroy T, Frécon G, Asquier-Khati A, Mateus D, Lecomte R, Rizkallah M, et al. 18F-FDG-Based Radiomics and Machine Learning. JACC Cardiovasc Imaging. 2023;16:951–61. https://doi.org/10.1016/j.jcmg.2023.01.020.
18. Swart LE, Gomes A, Scholtens AM, Sinha B, Tanis W, Lam MGEH, et al. Improving the diagnostic performance of 18F-Fluorodeoxyglucose Positron-Emission Tomography/Computed Tomography in Prosthetic Heart Valve endocarditis. Circulation. 2018;138:1412–27. https://doi.org/10.1161/CIRCULATIONAHA.118.035032.

Authors and Affiliations

D. ten Hove[1,2] · R. H. J. A. Slart[1,3] · A. W. J. M. Glaudemans[1] · D. F. Postma[4] · A. Gomes[2] · L. E. Swart[5] · W. Tanis[6] · P. P. van Geel[7] · G. Mecozzi[8] · R. P. J. Budde[9] · K. Mouridsen[1,10] · B. Sinha[2]

✉ D. ten Hove
d.ten.hove@umcg.nl

R. H. J. A. Slart
r.h.j.a.slart@umcg.nl

A. W. J. M. Glaudemans
a.w.j.m.glaudemans@umcg.nl

D. F. Postma
d.f.postma@umcg.nl

A. Gomes
a.gomes@umcg.nl

L. E. Swart
laurens.swart@gmail.com

W. Tanis
w.tanis@hagaziekenhuis.nl

P. P. van Geel
p.p.van.geel@umcg.nl

G. Mecozzi
g.mecozzi@umcg.nl

R. P. J. Budde
r.budde@erasmusmc.nl

K. Mouridsen
kim@cercare-medical.com

B. Sinha
b.sinha@umcg.nl

1 Department of Nuclear Medicine and Molecular Imaging, University of Groningen, University Medical Center Groningen, Medical Microbiology & Infection Prevention, Hanzeplein 1, Groningen 9713 GZ, The Netherlands

2 Department of Medical Microbiology and Infection Prevention, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

3 Biomedical Photonic Imaging group, Faculty of Science and Technology, University of Twente, Enschede, The Netherlands

4 Department of Internal Medicine and Infectious Diseases, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

5 Department of Cardiology, Spaarne Gasthuis, Haarlem, The Netherlands

6 Department of Cardiology, HagaZiekenhuis, The Hague, The Netherlands

7 Department of Cardiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

8 Department of Cardiothoracic Surgery, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

9 Department of Radiology and Nuclear Medicine, Erasmus University Medical Center, Rotterdam, The Netherlands

10 Department of Clinical Medicine, Center of Functionally Integrative Neuroscience, Aarhus University, Aarhus, Denmark