

Novel Applications of Machine Learning to Guide Diagnosis of Endocarditis

Axel Svenningsson



LUND
UNIVERSITY

Department of Automatic Control

BSc Thesis
TFRT-7672
ISSN 0280-5316

Department of Automatic Control
Lund University
Box 118
SE-221 00 LUND
Sweden

© 2025 Axel Svenningsson. All rights reserved.
Printed in Sweden by Tryckeriet i E-huset
Lund 2025

Abstract

Infective endocarditis (IE) is a severe disease affecting the inner lining of the heart. Its diverse symptoms complicate diagnosis and the current main diagnostic tool is the Duke criteria which classifies suspected cases as definite, possible or rejected IE. A substantial portion of cases falls into the ambiguous “possible” category and there is a need for a more precise diagnostic system.

This thesis aimed to investigate the usage of machine learning (ML) models, utilizing non-linear relationships, as a tool to determine whether a patient suffers from IE and extract what features are important in decision making. Multiple ML algorithm techniques were trained with a set of variables extracted from medical records. For target variable the opinion from a team consisting of medical specialists of multiple disciplines was used.

The results showed that all ML algorithms outperformed logistic regression, which was used as a linear model for comparison. The Duke criteria in turn outperformed the ML models in specificity and sensitivity. The two best performing models were based on tabular prior-data fitted network and random forest. The ML models largely utilized known predictors of IE but also identified new important variables including estimated glomerular flow rate, C-reactive peptide, no known focus and body mass index.

Acknowledgements

I would like to thank my supervisor, Mattias Ohlsson, professor at the Centre for Environmental and Climate Science at Lund University, for taking on this project. I truly appreciate his valuable guidance and support throughout this thesis. I also extend my gratitude to the helpful insights and feedback from my co-supervisor, Magnus Rasmussen, professor in Infectious Medicine at Lund University. Their support has been greatly appreciated.

Lastly, I would like to thank my examiner Bo Bernhardsson, professor at Department of Automatic Control, for giving me his time as well as useful feedback on the bachelor process.

Contents

1. Introduction	9
1.1 Background.....	9
1.2 Objectives and research questions.....	12
2. Methods	13
2.1 Data acquisition	13
2.2 Machine Learning Methods.....	13
2.3 Statistical Methods	19
3. Results	21
3.1 Participants	21
3.2 Descriptive data	21
3.3 Performance.....	24
3.4 Important features.....	28
4. Discussion and Conclusions	34
4.1 Discussion.....	34
4.2 Conclusion	37
4.3 Future Work.....	37
4.4 Critical Evaluation.....	38
5. Bibliography	40
6. Appendix	42
6.1 Appendix 1	42
6.2 Appendix 2	43
6.3 Appendix 3	45
6.4 Appendix 4	46

1. Introduction

1.1 Background

Infective Endocarditis

Infectious endocarditis (IE) is an infection of the inner lining of the heart, the endocardium. It is most commonly caused by bacteria after entering the blood stream through a wound, other infection or unknown source of entry. Every year 3-10 per 100 000 population is affected (1) and while being an uncommon disease, it can lead to dire consequences. Bacterial emboli originating from the heart can cause strokes or lack of oxygen in other parts of the body. Another feared complication is sepsis, meaning multiorgan failure caused by a misdirected immune response during an infection. These complications can lead to death and the overall mortality rate of IE is 25 % (1).

The most common presenting symptoms in patients that are later diagnosed with IE are fever, heart murmur upon auscultation with stethoscope and congestive heart failure resulting in shortness of breath and accumulation of fluids on the legs and in the lungs (2). These symptoms are not specific for IE and the diagnosis often offers a challenge for the treating physician. The most definite way to diagnose IE in practice is with a positive blood culture of a pathogen that typically causes IE in combination with echocardiography that shows sign of IE, such as vegetations (masses composed of fibrin, platelets, and infecting organisms), abscesses, mild to severe valve leakage or dehiscence of a valvar prosthesis (3). Echocardiography can be done over the chest, transthoracic (TTE), or through the esophagus, transesophageal (TEE). Other imaging modalities that can support the diagnosis are fludeoxyglucose-18 positron emission tomography in combination with computer tomography (FDG-PET/CT) and electrocardiography-gated cardiac computer tomography (cardiac-CT). However, IE can be present without imaging findings and positive blood cultures, and in these cases the clinical picture is guiding.

The treatment for IE is long-term high dosage intravenous antibiotic treatment, lasting several weeks, leading to drawn out hospital visits. Depending on vegetation size, disease control and pathogen species open heart surgery is sometimes considered the best option. The aggressive treatment plans implicate that treatment should not be given if not necessary, however, not treating when needed can cause inefficient therapies and worsening of the condition.

Endocarditis Teams

Because of the difficulty in diagnosing IE and deciding on treatment multidisciplinary endocarditis teams have been proposed to help physicians decide whether treatment is needed. These teams are composed of multiple specialties, including infectious disease specialists, cardiothoracic surgeons, cardiologists, and further disciplines when needed. The idea has shown promising results and when first evaluated in France 2009 it showed a reduced mortality rate from 18.5 % to 8.2 % (4). This was partly due to decisions to treat more patients which could result in over treating patients, but also because more patients with severe disease were accepted for surgery. Moreover, this approach promotes the standardization of antibiotic therapy and adherence to surgical indications in line with guideline recommendations.

In 2017 a multidisciplinary endocarditis team was formed in Lund, Sweden. The team gathers three times per week to discuss cases of suspected IE throughout southern Sweden and decide whether IE is present as well as which treatment benefits the patient the most. However, most cases of IE in Sweden, and globally, are not discussed by endocarditis teams. This places a high responsibility on the treating physician which relies on previous experience and existing treatment guidelines.

The Duke Criteria

For the past three decades the Duke criteria have been the dominating diagnostic system for IE. The details are described in Appendix 1, but here follows a brief description. It is based on clinical findings, symptoms and predisposing conditions being assigned major or minor points. These points are added together to label every episode of suspected IE as rejected, possible or definite IE. Major points are given for:

- Microbiological evidence: Positive blood cultures with typical IE microorganisms, such as *S. aureus*, viridans-group streptococci, *S. gallolyticus*, or community-acquired enterococci, from two separate blood culture sets, or three or more separate blood-cultures with atypical pathogens.
- Imaging evidence: Echocardiography or cardiac CT showing sign of IE, or FDG PET/CT indicating abnormal metabolic activity involving a valve or prosthetic material.

- Surgical finding: Evidence of IE observed during cardiac surgery if not major imaging nor pathologic criteria is fulfilled.

Minor points involve symptoms and predisposing conditions such as fever with a temperature above 38.0 °C, presence of a prosthetic heart valve, previous episode of IE, current intravenous drug use, pacemaker or other cardiac implantable electronic device (CIED), and vascular phenomena, such as emboli or intracranial bleeding. A diagnosis of definite IE requires one of the following: (a) two major criteria, (b) one major and three minor criteria, or (c) five minor criteria. If either one major and one minor criterion, or three minor criteria are met, the case is classified as possible IE.

A recognized limitation to the Duke criteria is that a large portion of cases are regarded as possible IE. This has been the case since the criteria were established, and updated versions of the Duke criteria have struggled to adequately address it. An evaluation of the most recent version by the International Society for Cardiovascular Infectious Diseases (ISCVI) published in 2023 (5) found that 41% of patients with *S. Aureus* found in blood cultures, the most common IE pathogen, were classified as possible IE while only 7.9% were regarded as definite IE (6).

Given these limitations, clinicians and future research would benefit from a new diagnostic system grounded in scientifically validated predictors and tested on an independent cohort. The goal of such a system should be to reduce the number of cases classified as possible IE while preserving high sensitivity.

Endocarditis and Machine Learning

The most frequent attempts at increasing the diagnostic accuracy for IE are to compare possible predictive factors in patients with suspected disease and match this with the final diagnosis. This is customary done by calculating the odds ratio for each factor which represents its strength of association with the outcome and assumes linearity between diagnostic indicators and the risk of having IE (7). Due to the complexity of diagnosing IE, linear models, such as linear regression, risk overlooking important variables, as interactions between factors may be masked by opposing effects.

In recent years machine learning (ML) has transformed multiple clinical disciplines. By training models on large datasets and imitating the way humans learn, they have managed to improve performance and accuracy in healthcare fields (8). In contrast to previous methods, ML provides promising non-linear approaches capable of detecting subtle patterns in patient data that traditional models might

miss. By training predictive algorithms on large, labeled datasets, it may be possible to enhance diagnostic accuracy, reduce the frequency of "possible IE" classifications, and support clinical teams in early risk stratification.

The use of ML to improve diagnostic tools for infective IE is still in its early stages. One proof-of-concept study involving 160 patients evaluated for suspected prosthetic valve endocarditis demonstrated that an ensemble of ML models outperformed the modified Duke criteria (9). Another study applied ML algorithms to identify IE in a cohort of 15 748 patients with *S. aureus* in blood cultures and found comparable accuracy to traditional risk scores (10). Key predictive features included age, prior IE, heart valve disease, and infection acquired out of hospital, which are all described as previously known risk factor.

1.2 Objectives and research questions

The primary objective of this thesis is to develop and evaluate ML models that can be used to determine whether a patient should be diagnosed with IE or not. To find new diagnostic tools the endocarditis team's decision will be used as target variable to diminish dependability on the Duke criteria and be open for new predictive factors. These models will be evaluated based on specificity and sensitivity to compare with simpler linear models such as logistic regression, as well as the Duke-ISCVID criteria. Once constructed, the models will be examined to help find new important features when diagnosing IE.

The specific research questions are:

- How do ML models, trained and tested on separate data, with the endocarditis team's opinion as target variable, compare to linear models such as logistic regression, as well as to the Duke-ISCVID criteria, in the context of diagnosing IE?
- Which ML models have the highest sensitivity and specificity for diagnosing IE, and is thus considered the best approach for future research with larger datasets?
- By examining what features are important in decision making and what causes the model predictions to choose the correct or incorrect treatment plan, what can we learn about the ML models?

2. Methods

2.1 Data acquisition

The study included patients referred to the multidisciplinary endocarditis team in Lund between 2021 and 2023. A retrospective cohort was formed from this group. Clinical presentation, imaging results, and microbiological findings were collected from medical records. The data was collected by the author of this thesis and one more master's degree student. Data selection was guided by the Duke-ISCVID criteria, which were expanded with additional variables, including the study's outcome variable, the endocarditis team's decision. A full list of variables and their definitions can be found in Appendix 2. Each episode of IE was considered complete when antibiotic treatment was concluded, as determined by the treating physician. Consequently, a single patient could have multiple distinct IE episodes.

Exclusion criteria were as follows:

- Referrals from outside Skåne due to limited access to medical records.
- Missing personal identification numbers, preventing access to medical records.
- Referrals solely for removal of cardiac implant.
- Patients under 18 years of age at the time of referral.
- Patients who died before being evaluated by the endocarditis team or for whom care was redirected to palliation.
- Referrals unrelated to suspected IE.

2.2 Machine Learning Methods

A total of 54 input variables were used and the endocarditis team's decision to treat as IE or dismiss the diagnosis was used as target variable. Binary variables were encoded as 0 and 1, and continuous variables were normalized to zero mean and

standard deviation of one. Seven machine learning techniques were used to train models, resulting in seven different models. For training, validation and testing the python library scikit-learn was used (11).

Logistic Regression

Logistic regression (LR) was used as a linear baseline model for comparison with other machine learning techniques. In this model, each feature is assigned a coefficient, which is multiplied by the corresponding input value. The resulting products are summed and passed through a sigmoid function, resulting in an output between 0 and 1 that represents the probability for each episode of suspected IE. Different classification thresholds can then be applied to balance the trade-off between sensitivity and specificity. The model was trained using stochastic gradient descent, resulting in the final weights.

LR can be used without hyperparameters, but to increase generalization, regularization was added in the form of elasticnet, which is a combination of L1 and L2 regularization, and the ratio between these can be adjusted. L2 regularization discourages large weights by penalizing their squared values, promoting smoother models, while L1 regularization penalizes the absolute values of weights, encouraging sparsity by driving some weights to zero.

Multilayer Perceptrons

Multilayer perceptrons (MLPs) are among the simplest forms of artificial neural networks. MLPs make predictions by passing input data through multiple layers of interconnected nodes. Each input feature is connected to every node in the first hidden layer via a weight, which is multiplied by the feature value. The weighted sum at each node is then passed through an activation function. This output is then used as input for the next layer, again combined with weights, and passed through another activation function. The number of hidden layers as well as node in each layer can be tailored to perform optimal on the given task. In classification problems, the final layer typically consists of a single node, which applies the sigmoid function to the weighted sum from the last hidden layer, producing a probability between 0 and 1.

To improve generalizability L2 regularization was used. The strength of the regularization is proportional to the constant alpha. As activation function for the hidden layers, multiple choices are available. Since we used binary variables together with continuous variables normalized around 0 it aligns well with the shape

of tanh, which outputs values between -1 and 1. This helps the neural network learn symmetrical patterns more efficiently.

Training an MLP for our classification task was done using supervised learning in an iterative process involving three main steps:

1. **Forward pass:** The input data was passed through each layer resulting in a prediction from the output layer by the sigmoid function.
2. **Loss computation:** The output was compared to the true label and the loss was calculated using the binary cross-entropy function.
3. **Backward pass:** The loss function was used to update the network's weights through backpropagation. This was done through different methods such as Adam or stochastic gradient descent (SGD).

The extent to which the weights were updated in each cycle was controlled by the learning rate, which was tuned to optimize training performance. Before calculating the loss and updating the weights, multiple predictions were made to create a more accurate approximation of the gradient. The number of predictions that one update was based on is called batch size which also was varied to optimize training.

For our task, two types of MLPs were used. Firstly, a single multilayered perceptron with one hidden layer where hyperparameters were tuned to optimize training. Deep learning networks with multiple layers were tried shortly without success. Secondly, an ensemble of 10 MLPs were used (MLPE). Ensembles combine multiple models to reduce variance and improve overall prediction accuracy and robustness. The models were trained through bagging, which is a process based on bootstrapping.

Support Vector Machines

Support vector machines (SVM) are similar to logistic regression in that they aim to find a decision boundary, but instead of using a standard loss function they utilize the hinge loss function. The hinge loss function penalizes not only incorrect predictions but also correct predictions within a certain margin. This causes the model to strive for a separation of the classes with the widest possible margin.

To allow for more flexible decision boundaries SVM can use kernel functions. The simplest is the linear kernel which assumes data to be linearly separable in the original feature space. Alternatively, the radial basis function (RBF) can be used as kernel which combines variables non-linearly, enabling decision boundaries to be curved and thus more flexible.

To increase generalizability regularization was used. This is done with the constant C which determines the margin of penalized predictions near the decision margin. Decreasing C leads to a larger margin, decreasing overfitting.

Tree-Based Learning

Tree-based learning uses criteria to divide cases by certain features. The idea is to form a flow chart that separates cases by assigning thresholds or conditions. A group of inputs that fulfills a certain condition is then divided again by new conditions. This is done multiple times until a certain target variable has been identified.

Tree-based learning can be performed with a single tree but there are advantages in combining multiple trees in an ensemble. By allowing each tree to deal with different parts of the problem variance can be reduced, assuming errors are not positively correlated.

Creating an ensemble of decision trees can be done in multiple ways. One technique is called random forest (RF). It utilizes perturbed trees that are trained on bootstrapped datasets. The trees can be perturbed in multiple ways to reduce variance and prevent overfitting. For instance, only a limited amount of input variables can be allowed for training, different methods can be used to calculate the impurity which represents the homogeneity of the next node after each split, or a randomness can be introduced when deciding on thresholds.

Another method is called boosting which also creates an ensemble but in contrast utilizes sequential learning, where each new model is trained to correct the errors of the previous one. This creates a more flexible model which aims to reduce bias. The contribution from each tree on the classification task is given by a weight. In this thesis we used a python library called XGBoost (12), which is a highly efficient and scalable implementation of gradient boosting.

There are multiple ways to regularize ensembles of decision trees. One can limit the depth of each tree, demand a minimum number of samples required to split a node, decide the number of trees in the model, or limit the number of features to be considered for each tree or for each split. In the case of boosting algorithms, one can determine the fraction of samples that are used in each boosting round and adjust learning rate, which controls the contribution from each tree to the final mode. One can also state a minimum sum of instance weights, or sum of Hessians (i.e. second derivatives of the loss function) required for a child node. This constraint helps prevent overfitting by disregarding splits that involve too few or low-importance samples.

Tabular Prior-Data Fitted Network

Tabular Prior-Data Fitted Network (TABPFN) is a pre-trained foundation model, meaning that it has been trained on a vast number of tabular datasets to be able to detect patterns in new, previously unseen, data (13). Tabular datasets are, like ours, structured data organized in rows and columns, usually containing one target column. The idea is that similarly to pictures containing recurring features, such as edges, shapes, and textures, tabular data also share certain statistical and structural patterns that have been learned from both real and synthetic datasets. This allows TABPFN to generalize well to new, previously unseen datasets in smaller samples without having to relearn the shared patterns.

While the model is pretrained to detect typical patterns in tabular datasets, a training set is needed for the in-context learning. In contrast to the other models, there are no hyperparameters to tune.

K-Fold Validation and Testing

Because of the relatively small sample, K-fold was used for validation and testing, see figure 1. This was done by splitting the dataset in five equal parts using stratification. Four of the parts were used to train the data and the fifth was used to measure generalization on unseen data. Within the training data the episodes were split into five parts further, again with stratification, to find optimal hyperparameters. This was done by training a wide range of hyperparameters on four of the splits and validating them on the fifth. The trial and error were done in all five combinations by excluding a unique set each time, and the entire validation process was repeated 10 times to make sure the best model was found. For each training fold, 250 random combinations of hyperparameter sets were tried. The range of tested hyperparameters is presented in table 1. The hyperparameters that resulted in the model that performed best on the validation sets were used to train the entire training set (80 % of the entire dataset) which was then tested on the test set (20 % of the entire dataset). This entire process was then repeated five times by excluding a unique set each time, and the entire process was done twice, resulting in 10 final models for each ML technique.

To quantify the effectiveness of the methods the area under curve in the receiver operating characteristic curve (ROC-AUC) was calculated.

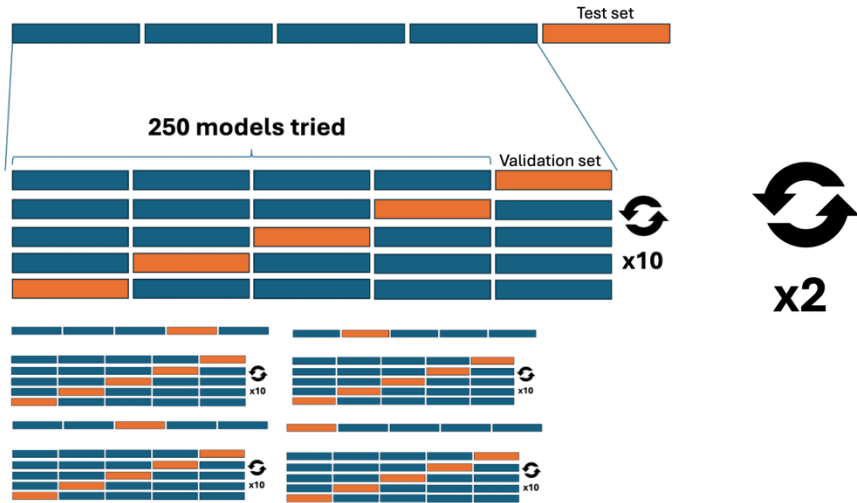


Figure 1. Visual of how the K-fold validation was performed. 80% of data was used to train the data and 20% was used to test performance. This was done for all five folds, twice. K-fold was in turn used to find optimal hyperparameters by trying 250 random combinations of hyperparameters on four fifth of the training data and validating on the fifth. This was repeated for all five parts and cycled ten times.

Table 1. The range of hyperparameters tried for each machine learning model. Log-uniform describes a logarithmic uniform distribution. Parenthesis is used to describe ranges. Randint gives a random integer in the given range. Maximum features describes how many features are considered when splitting a node as a function of all features (square root or log base 2). Subsample describes the fraction of the data that is randomly used for each boosting round. Colsample by tree tells the fraction of features used for each tree. Tanh is the hyperbolic tangent function.

<i>Model</i>	<i>Hyperparameters and ranges</i>
<i>Logistic regression (LR)</i>	$C \sim \text{log-uniform}(10^{-2}, 10)$ penalty: elasticnet L1 ratio $\sim \text{uniform}(0,1)$
<i>Support vector machine (SVM)</i>	kernel = linear and RBF $C \sim \text{log-uniform}(10^{-5}, 10^2)$ hidden layer sizes = 2, 3, 5, 7, 10
<i>Multilayer perceptron (MLP)</i>	$\alpha \sim \text{log-uniform}(10^{-2}, 1)$ activation function = tanh batch size = 16, 32, and 64 learning rate = adaptive initial learning rate = 0.01, 0.02, and 0.05

<i>Multilayer perceptron ensemble (MLPE)</i>	solver = SGD and adam number of estimators = 10 base MLP with same as above
<i>Random forest (RF)</i>	max depth ~ randint(3, 10) minimum samples split ~ randint(2, 5) number of estimators = 50, 100, 200, 500, and 2000 maximum features = sqrt and log2
<i>XGBoost (XGB)</i>	maximum depth = 2, 3, 4, 5, 6, 7, and 8 minimum child weight = 1, 2, 5, and 10 subsample = 0.5, 0.7, and 1.0 colsample by tree = 0.25, 0.5, 0.75, and 1.0 learning rate = 0.005, 0.01, 0.05, 0.1, 0.2, and 0.5 number of estimators = 50, 100, 300, and 500
<i>Tabular Prior-Data Fitted Network (TABPFN)</i>	no hyperparameters

2.3 Statistical Methods

Determining Significance

For descriptive data the Chi-2 and Mann-Whitney U tests were used to calculate p-values for binary and continuous variables respectively. To determine significance between logistic regression and the more advanced ML techniques, a modified t-test based on repeated k-fold cross validation was used (14). This was needed as performance estimates across folds and repetitions are not independent. Level of significance was defined as 0.05.

Shapley Additive Explanation

Shapley additive explanations (SHAP) is a game theoretic approach to explaining the output of an ML model (15). SHAP assigns each feature an importance value, which represents its contribution to an individual prediction. The feature importance varies for each prediction a model makes and the sum of all feature's feature importances results in the model's prediction. This method allowed for insight in what factors had the most impact on a specific case and was used to find limitations to a model by evaluating wrong predictions. By calculating the mean feature importance, the most important overall features were also identified for each ML

technique. This was done by averaging the feature importance across all 10 final models in k-fold testing. As a point of comparison, feature importance using the built-in permutation importance from the RF model was also used.

Ethical Considerations

An ethical application was accepted by the local ethical committee (permit 2017/1003). As this is a retrospective study relying on data regarding clinical presentation and relevant test results, access to medical records was necessary. This highlights the importance of maintaining confidentiality and maintaining patient privacy. This comes with the responsibility to ensure that no information is disclosed to unauthorized individuals, as doing so would not only breach data security and violate patient rights, but also risk undermining public trust in medical research.

As the study was retrospective, explicit consent was not asked for. This approach may raise ethical concerns, such as patients feeling that their personal experiences were used without permission. However, the potential benefits of developing improved diagnostic tools for this patient group justify the method. Requiring explicit consent would have introduced significant selection bias and posed a large logistic challenge in tracing and informing all patients.

IE is a life-threatening condition, and improved diagnostic and therapeutic strategies have the potential to benefit current and future patients. Recurring IE remains a clinical issue, further emphasizing the need for advancements, directly for the subjects.

All patient data were handled in compliance with the Swedish Patient Data Act (2008:355).

3. Results

3.1 Participants

A total of 585 episodes of IE were referred to the endocarditis team in Lund between 2021 and 2023, see figure 2. Seven episodes were excluded due to no suspicion of IE. Two patients were diseased, or motive of care was changed to palliation before a decision was made. 141 episodes were excluded due to referral from outside of Skåne. 11 patients had an age below 18 years and 19 episodes only regarded removal of CIED. The remaining 405 episodes constituted the study cohort. Of these, 272 episodes were regarded as IE by the team and 133 were rejected.

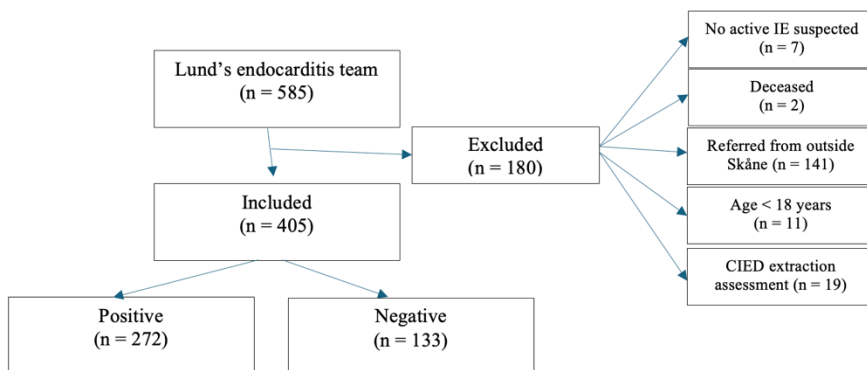


Figure 2. The patient cohort of episodes of IE referred to the endocarditis team 2021-2023. Reasons for exclusion is shown to the right.

3.2 Descriptive data

Table 2 presents the entire list of variables collected and the occurrence of these in all suspected, confirmed, and rejected episodes of IE respectively. For binary variables the number of occurrences is shown and for continuous variables the median is presented. The corresponding p-value is calculated between the groups of IE and no IE. For binary variables the Chi squared test was used while the Mann-Whitney U test was used for continuous variables. The level of significance was determined at 0.05. It can be noticed that 17 variables differed significantly between

the IE and no IE groups. The largest discrepancy between these groups was for vegetation on TEE/TTE which occurred in 73 % of episodes if IE but only 20 % of episodes where IE was rejected. For continuous variables, only number of positive bottles was significant where the median for the IE and rejected group was 4 and 2 respectively.

Table 2. All variables collected for all episodes, episodes of IE and episodes of no IE are shown respectively. For binary variables the number of occurrences is shown, and percentage of occurrence is presented in parenthesis. For continuous variables the median is shown with interquartile range in parenthesis. p-values between episodes of IE and no IE are shown for each binary variable calculated with the Chi-2 test and continuous variables with the Mann-Whitney U test. Level of significance was 0.05.

Variable	All episodes (n=405)	IE (n=272)	No IE (n=133)	p-value
Predisposing factors				
Age / years	73 (63-81)	74 (61-81)	72 (63-80)	0.95
Sex (female)	106 (26%)	71 (26%)	35 (26%)	1.0
BMI	25 (23-30)	25 (23-29)	25 (23-30)	0.81
Native valve disease	111 (27%)	90 (33%)	21 (15%)	< 0.001
CIED	104 (26%)	63 (23%)	41 (31%)	0.10
Previous IE	58 (14%)	46 (17%)	12 (9.0%)	0.039
Intravenous drug use	26 (6.4%)	23 (8.5%)	3 (2.3%)	0.030
Prosthesis	119 (29%)	79 (39%)	40 (30%)	0.10
Bio prosthesis	123 (30%)	91 (33%)	32 (24%)	0.069
Mechanical prosthesis	19 (4.7%)	13 (4.8%)	6 (4.6%)	1.0
Baseline eGFR	71 (58-87)	70 (58-88)	72 (57-85)	0.80
Clinical findings				
Vegetation on TEE/TTE	225 (56%)	198 (73%)	27 (20%)	< 0.001
Abscess on TEE/TTE	29 (7.2%)	26 (9.6%)	3 (2.3%)	0.013
Leakage on TEE/TTE	44 (11%)	41 (15%)	3 (2.3%)	< 0.001
Finding on CT	14 (3.5%)	12 (4.4%)	2 (1.5%)	0.22
Finding on FDG-PET/CT	28 (6.9%)	24 (8.8%)	4 (3.0%)	0.050
Heart murmur	227 (56%)	169 (62%)	58 (44%)	< 0.001
Seven days symptoms	152 (38%)	113 (42%)	39 (29%)	0.015
Emboli	103 (25%)	83 (31%)	20 (15%)	0.0012
Days of fever	1 (0-2)	1 (0-2)	1 (0-2)	0.57

Maximum CRP	142 (83-230)	144 (83-231)	141 (86-227)	0.46
CRP on day five	60 (30-120)	62 (33-119)	53 (29-119)	0.39
Kidney injury	78 (19%)	55 (20%)	23 (17%)	0.57
Acute eGFR	50 (25-70)	51 (25-71)	47 (23-68)	0.37
Infectious focus				
Focus: spondylodiscitis	39 (9.6%)	30 (11%)	9 (6.8%)	0.24
Focus: none	305 (75%)	212 (80%)	93 (70%)	0.10
Focus: joint or bone	40 (9.9%)	28 (10%)	12 (9.0%)	0.82
Focus: lung	21 (5.2%)	13 (4.8%)	8 (6.0%)	0.77
Focus: urinary tract	14 (3.5%)	6 (2.2%)	8 (6.0%)	0.09
Focus: gastrointestinal	7 (1.7%)	2 (0.74%)	5 (3.8%)	0.074
Focus: teeth	6 (1.5%)	5 (1.8%)	1 (0.75%)	0.68
Focus: muscle	5 (1.2%)	4 (1.5%)	1 (0.75%)	0.89
Focus: other	7 (1.7%)	2 (0.74%)	5 (3.8%)	0.074
Blood cultures				
Number of positive bottles	2 (2-4)	4 (2-4)	2 (0-4)	< 0.001
Polymicrobial bacteremia	21 (5.2%)	12 (4.4%)	9 (6.8%)	0.44
Negative culture	54 (13%)	18 (6.6%)	36 (26%)	< 0.001
Persistent bacteremia	55 (14%)	44 (16%)	11 (8.2%)	0.043
<i>S. aureus</i>	108 (27%)	69 (25%)	39 (29%)	0.47
<i>E. faecalis</i>	47 (12%)	38 (14%)	9 (6.8%)	0.050
<i>S. mitis</i>	36 (8.9%)	29 (11%)	7 (5.3%)	0.11
<i>S. sanguinis/gordonii</i>	30 (7.4%)	26 (9.6%)	4 (3.0%)	0.031
<i>S. bovis</i> complex	19 (4.7%)	15 (5.5%)	4 (3.0%)	0.38
<i>S. dysgalactiae</i>	18 (4.4%)	12 (4.4%)	6 (4.5%)	1.0
KNS	15 (3.7%)	10 (3.7%)	5 (3.8%)	1.0
Non-HACEK Gram-negative Enterobacterales	10 (2.5%)	7 (2.6%)	3 (2.3%)	1.0
<i>S. agalactiae</i>	8 (2.0%)	5 (1.8%)	3 (2.3%)	1.0
<i>S. salivarius</i>	6 (1.5%)	6 (2.2%)	0 (0.0%)	0.20
<i>S. mutans</i>	5 (1.2%)	5 (1.8%)	0 (0.0%)	0.27
Corynebacterium spec.	5 (1.2%)	3 (1.1%)	2 (1.5%)	1.0
<i>S. lugdunensis</i>	5 (1.2%)	3 (1.1%)	2 (1.5%)	1.0
<i>S. anginosus</i>	5 (1.2%)	2 (0.74%)	3 (2.3%)	0.41
Other species	35 (8.6%)	24 (8.8%)	11 (8.3%)	1.0
Acquisition type				
Community acquired	195 (48%)	150 (55%)	45 (34%)	< 0.001
Healthcare associated	160 (40%)	97 (36%)	63 (47%)	0.031
Nosocomial	49 (12%)	25 (92%)	24 (18%)	0.016

3.3 Performance

The performance measured as ROC-AUC of the ML models including logistic regression is visualized in figure 3 and table 3. All more advanced ML models outperformed logistic regression, and the difference was on average 4.5 %. The discrepancy was significant for MLP, MLPE, RF, XGB and TABPFN. SVM outperformed logistic regression but did not reach significance. The best hyperparameters for each model can be seen in appendix 3.

The average ROC curve over all 10 folds for each model can be seen in figure 4. The sensitivity and specificity of Duke-ISCVID is presented with one point. While outperforming logistic regression, the ML models did not manage to perform as well as Duke-ISCVID in the single point. For interest and further discussion an attempt to imitate the Duke criteria was made by training the data with the Duke classification as target variable, which is presented in appendix 4. This was done for two ML models using MLP and RF. Both models managed to reach a ROC-AUC of 0.96.

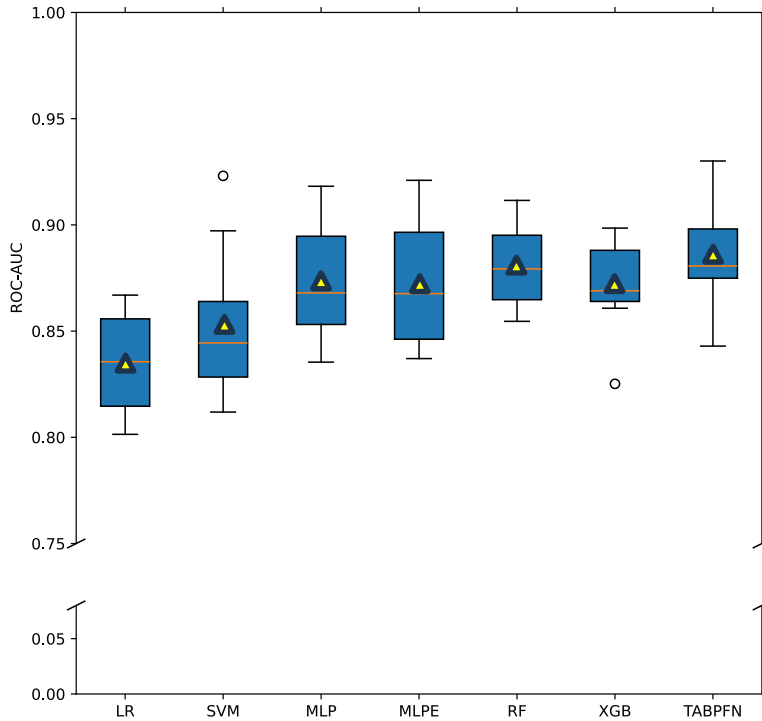


Figure 3. Area under curve for the receiver operating characteristic curve (ROC-AUC) from 10 measurements of each model on previously unseen test data. Explanations: Triangle - mean, Line - median. Box - interquartile range. Lines: quartile 0.5 to 1.0 and quartile 3.0 to 3.5, Circles: outliers. Abbreviations: LR - stepwise forward multivariate logistic regression. SVM - support vector machine. MLP - multilayer perceptron. MLPE - multilayer perceptron ensemble. RF - random forest. XGB - XGBoost. TABPFN - tabular prior-data fitted network. ROC-AUC - area under curve for the receiver operating characteristic curve.

Table 3. Mean, median and standard deviation from 10 measurements of area under curve for the receiver operating characteristic curve (ROC-AUC). p-values are based on comparison with logistic regression, calculated as described in Bouckaert et al (2004). Level of significance was 0.05. Abbreviations: ROC-AUC - area under curve for the receiver operating characteristic curve. LR – logistic regression. SVM – support vector machine. MLP(E) – multilayer perceptron (ensemble). RF – random forest. XGB – XGBoost. TABPFN - Tabular prior-data fitted network.

Model	Mean ROC-AUC (n = 10)	Median ROC-AUC	Std ROC-AUC	p-value (LR)
LR	0.835	0.836	0.0232	-
SVM	0.853	0.844	0.0331	0.31
MLP	0.874	0.868	0.0283	0.010
MLPE	0.872	0.868	0.0288	0.014
RF	0.880	0.879	0.0183	< 0.001
XGB	0.871	0.869	0.0204	0.016
TABPFN	0.885	0.881	0.0234	0.0035

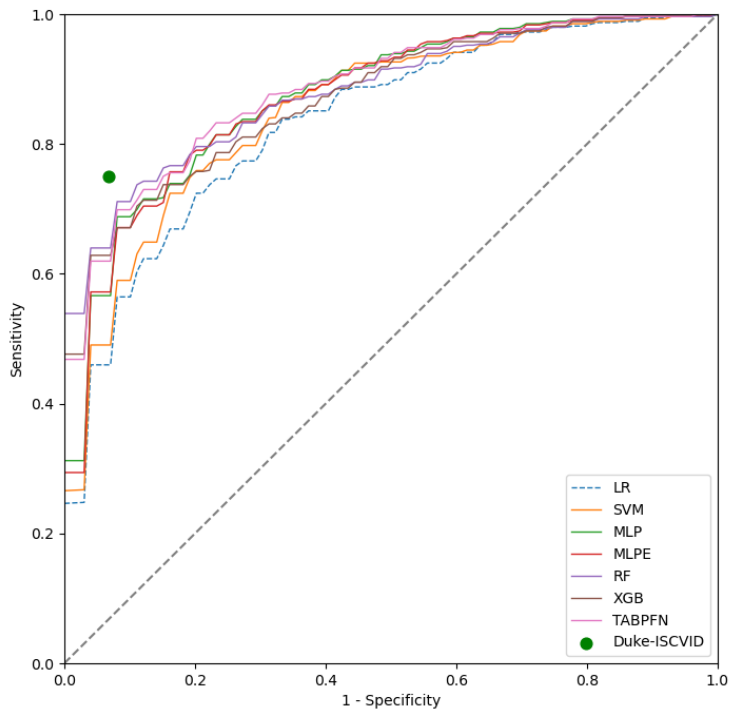


Figure 4. Mean receiver operating characteristic (ROC) curve for 10 models of each algorithm. Duke-ISCVID is based on one single measurement. Abbreviation: SVM – support vector machine. MLP – multilayer perceptron. MLPE – multilayer perceptron ensemble. RF – random forest. XGB – XGBoost. TABPFN - Tabular prior-data fitted network. LR - stepwise forward multivariate logistic regression. ISCVID - International Society for Cardiovascular Infectious Diseases.

3.4 Important features

Figure 5 shows the 15 highest ranked variables in feature importance for the models MLP, MLPE, RF, XGB and TABPFN. For RF the built-in permutation explainer is presented as well. Vegetation on TEE/TTE was the most important feature in all models analyzed. The second ranked variable in importance was acute estimated glomerular flow rate (eGFR), which is an estimate of acute kidney function, in MLP and MLPE, and blood culture related variables in RF, XGB and TABPFN, such as number of positive bottles and negative culture. Other variables that ranked highly in multiple models were native valve disease, emboli, maximum C-reactive protein (CRP), which is an inflammation marker, and CRP on day five.

Figure 6 shows bee swarm plots for a selected group of features in the models MLP, MLPE, RF, XGB and TABPFN. In contrary to feature importance, these plots enable interpretation regarding in which direction certain feature values impact the model. The plots are based on SHAP values calculated on one model for each ML technique that was trained on the entire dataset. Visually, all analyzed features were associated with an increased risk of having IE, except baseline eGFR (except for XGB), CRP on day five and negative culture. For the tree-based models (RF and XGB) there is a greater mixture between the groups for baseline eGFR, number of positive bottles and maximum CRP.

Three examples of predictions can be visualized in figure 7. Firstly, a wrong prediction by the MLP model, relying too heavily on vegetation on TEE/TTE, is shown. Further, two correct predictions by RF are shown. One prediction is positive for IE despite negative echocardiography and the second is a rejected episode with evidence of vegetation shown on TEE/TTE. To compare the impact of echocardiography on the models, table 4 presents how many predictions with and without vegetation on TEE/TTE are declared positive and negative for IE. For the episodes with evidence of IE on TEE/TTE, the models MLP, MLPE, XGB and TABPFN assigned zero to one as negative while RF predicted that 11 are negative. The endocarditis team's decision was that 12 of the echocardiography positive episodes were not suffering from IE. For the episodes without vegetation on TEE/TTE the predictions were more heterogenous.

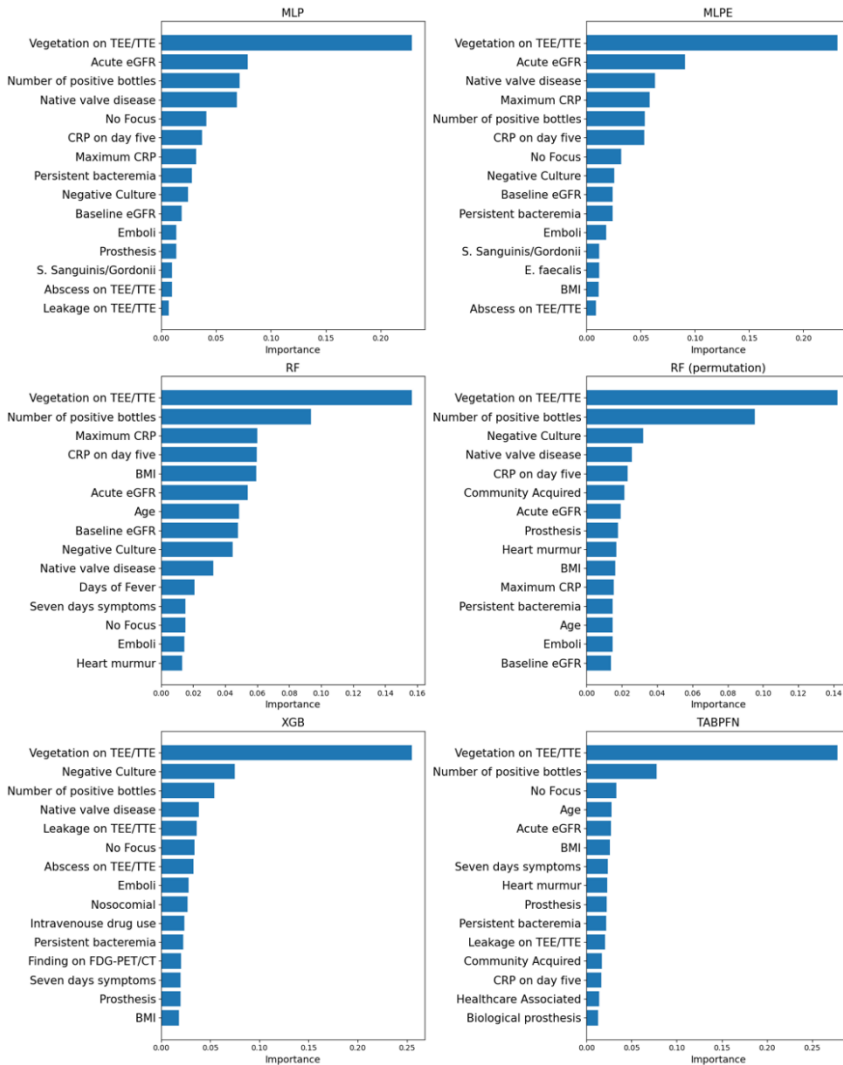
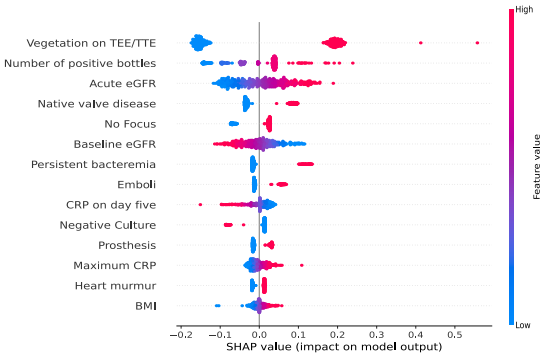
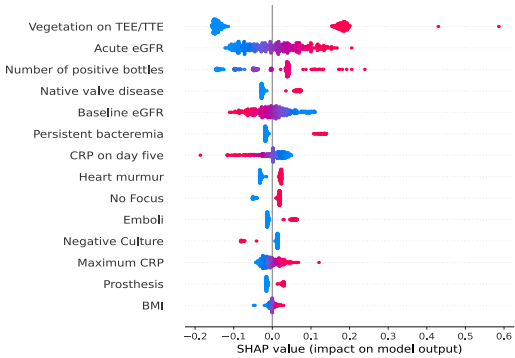


Figure 5. The highest ranked variable importances from five of the machine learning models according to SHAP. For random forest the built-in permutation explainer was used as well. Variable importance is a measurement representing how much each input variable contributes to a model's predictions. Abbreviations: MLP(E) – multilayer perceptron (ensemble). RF – random forest. XGB – XGBoost. TABPFN - Tabular prior-data fitted network. TEE/TTE - transesophageal or transthoracic echocardiogram. eGFR – estimated glomerular flow rate. CRP – C-reactive protein. BMI – body mass index.

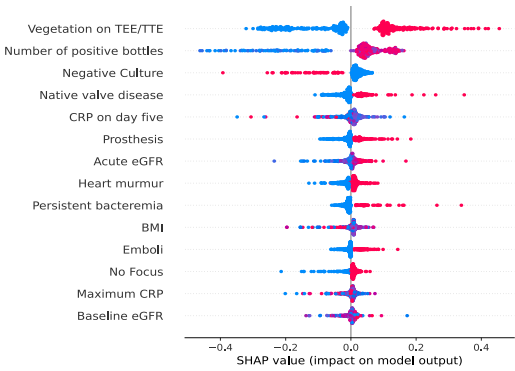
MLP



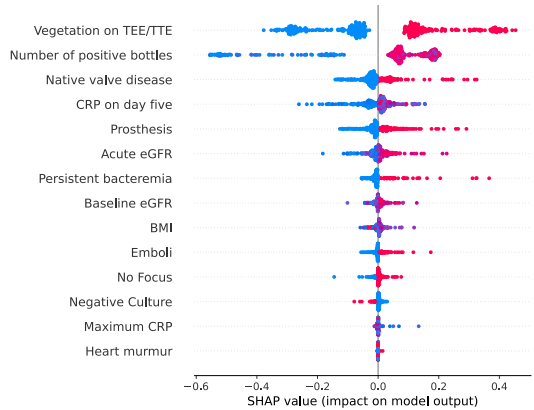
MLPE



RF



XGB



TABPFN

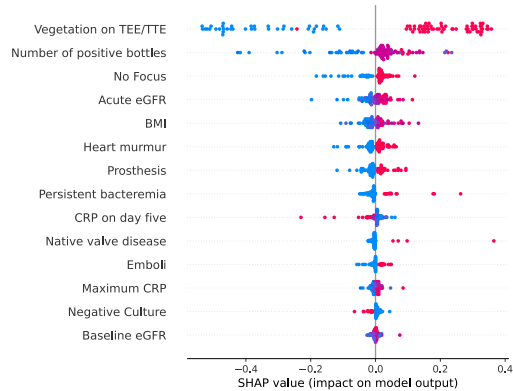


Figure 6. Bee swarm plots showing selected variable values as color and SHAP values on the x-axis. SHAP values explain to what extent each variable contributed to a specific prediction made by a machine learning model and in which direction (increased or decreased risk of infective endocarditis). For TABPFN a limited sample was used to decrease calculation times. Abbreviations: MLP – multilayer perceptron. MLPE – multilayer perceptron ensemble. RF – random forest. XGB – XGBoost. TABPFN - Tabular prior-data fitted network.

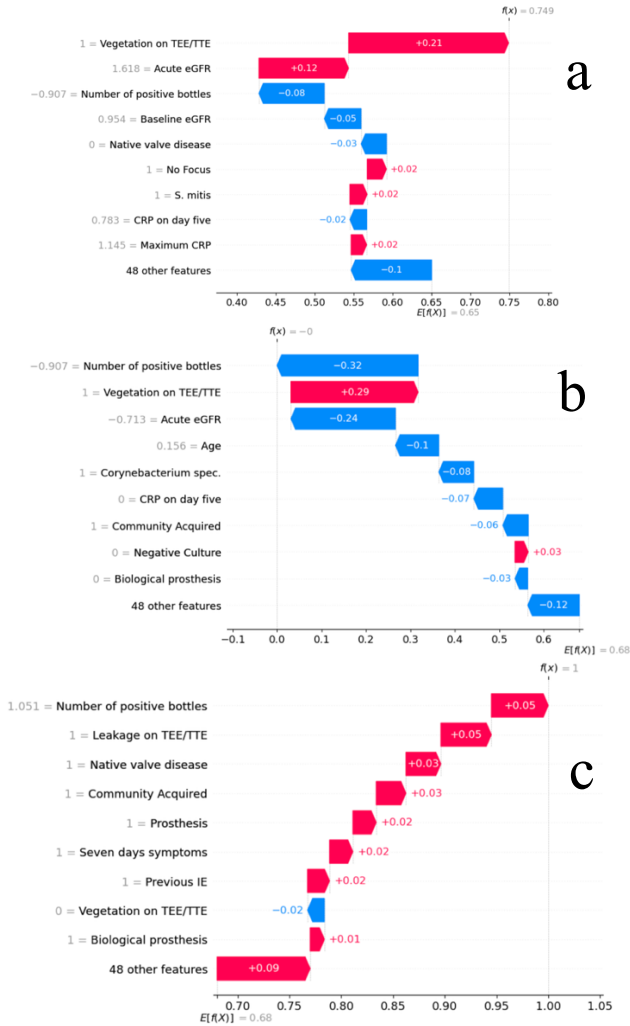


Figure 7. Three examples of predictions with different models. The variable value is shown to the right and its contribution to the prediction by the red or blue arrow. $f(X)$ is the prediction representing a probability and $E[f(X)]$ is the mean prediction for all samples. a: MLP predicting wrong (positive) largely due to vegetation on TEE/TTE. b: RF predicting correct (negative) despite vegetation on TEE/TTE. c: RF predicting correct (positive) despite no vegetation on TEE/TTE.

Table 4. The predictions for five of the machine learning models divided in vegetation on TEE/TTE and no vegetation on TEE/TTE. The opinions of the endocarditis team is shown on the last row. Abbreviations: pos – positive predictions for IE. neg – negative predictions for IE.

Modell	Vegetation (pos/neg)	No vegetation (pos/neg)
MLP	188/1	97/119
MLPE	188/1	104/112
RF	178/11	102/114
XGB	189/0	132/84
TABPFN	188/1	105/111
<i>Endocarditis team</i>	<i>177/12</i>	<i>95/121</i>

4. Discussion and Conclusions

4.1 Discussion

Performance

One major finding in this thesis is that most of the ML models managed to significantly outperform the simpler linear logistic regression model on previously unseen data. This indicates that advanced models are more capable of predicting whether a patient suffers from IE and that future research should aim to further investigate interfactorial relationships. Research aimed at finding linear relationships between diagnostic criteria and the diagnosis is necessary but risk missing these relationships.

Multiple models managed to outperform the logistic regression model, and the top performer was the TABPFN model. TABPFN has the advantage of being pre-trained on numerous tabular datasets which makes it ideal for small samples like the one we are working with. It diminished the need for tuning of hyperparameters and has a head start regarding recurring patterns that are often present in tabular data. The second-best model was RF. For small datasets, meaning less than 10 000 samples, decision forests have been shown to generally perform better on tabular data than MLPs (16), which is consistent with our results. XGB, MLP and MLPE all performed on an equal but lower level and probably demands larger datasets to excel through finer tuning. Early in the process deep neural networks with more layers were tried without success. This is likely also due to the small sample size inhibiting it from generalizing well. Another reason could be that the underlying relationships in the data can be captured by simpler models and that the deep networks introduce unnecessary complexity, leading to overfitting. The fact that SVM performed similar to logistic regression further supports that while there appears to be complexity and non-linear relationships the predictive signal is to some degree linearly structured.

The results showed that the Duke-ISCVID criteria better managed to imitate the endocarditis team than the logistic regression and ML models at the single point. This can be for multiple reasons. One explanation is that the predictive signal is too

complex to learn on this limited sample size. For instance, learning which bacteria are typical and atypical causers of IE is included in the Duke-ISCVID criteria after decades of experience and research, but might not be deductible from this dataset as some species are only present in five episodes.

To investigate this further we trained the ML models with the Duke-ISCVID criteria as target variable. The fact that the ML models managed to almost with certainty imitate the criteria with a ROC-AUC of 0.96 speaks against this. Another reason could be an unpredictability of the endocarditis team. This could be for two reasons. Firstly, there could be variables upon which the team bases their decision that are not included among the input variables. This could be for instance an altered immune system due to cancer treatment or another underlying medical condition, or a more complex view of the echocardiography which is not always as simple as showing clear signs of IE or not. Secondly, it could be caused by some randomness being introduced to the decision by the human variable. Either a single physician could be inconsistent when deciding on treatment, or the decision could be somewhat dependent on which physician is present, as there is a rotation of professionals from week to week.

One benefit of using ML as opposed to the Duke-ISCVID criteria is that the sensitivity and specificity can be adjusted. This can be used either as a screening method or to rule out IE before consulting the endocarditis team. With a specificity of near 100 % there is a sensitivity of around 50 % for most of the ML models which can be utilized to decide to treat without involving the team. On the contrary, a sensitivity of near 100 % gives a specificity of around 20 % which can be used to rule out the diagnosis immediately.

Important features

Feature importance identified vegetation on TEE/TTE as a dominating variable when deciding whether a patient has IE or not, which is in line with previous research (17). Ranked after that is a mixture of known predictive factors and new ones. Blood culture dependent variables, native valve disease, emboli, prosthesis, loci of other infectious focus, and heart murmur are among the highest ranked variables that are known diagnostic predictors. However, lesser-known predictive factors were ranked higher than expected, including acute eGFR, baseline eGFR, acute CRP, CRP on day five, body mass index (BMI), and lack of known focus. These variables are not primarily considered interesting among physicians when differentiating IE from other similar conditions and offer new interesting potential

tools. It is also clear that the built-in permutation importance from RF gives similar results which is expected.

Interestingly the bee swarm plots hint that the dynamic change in both eGFR and CRP are important predictors of IE. A higher acute eGFR, representing better kidney function during admission, was linked with an increased risk of having IE. In contrast, a higher habitual eGFR, marking superior kidney function before the infection, was associated with a lower risk of having IE. A possible explanation for this is that patients with more severe illness, and thus more likely to suffer from kidney injury, are to a greater extent referred to the endocarditis team, as diagnostic oversight would have more dire consequences. The other reason could be that IE does not affect the kidneys to the same extent as similar differential diagnoses, which would contradict previous research on the subject.

For CRP, the maximum value during the first five days of admission was positively correlated with having IE, while an elevated value on day five predicted a lower probability of having IE. This suggests that IE is characterized by a quick spike in CRP followed by a decline within the first five days. Another possibility is that patients are referred to the endocarditis team already receiving IE treatment with the decline in CRP representing an effective regimen.

Lastly, the bee swarm plots showed an increased risk of having IE for patients with a lower BMI, which has been observed in one previous study (18), but demands further research. The direction in which a variable affected the prediction was consistent in all ML models, with the exception of XGB having reverse correlation with baseline eGFR.

Vegetation on TEE/TTE greatly impacted all models. They did not manage to identify episodes with evidence of IE on echocardiography but where the diagnosis was rejected by the endocarditis team. The exception was RF and SHAP values indicate that it used number of positive bottles and acute eGFR to compensate.

The shape of the bee swarm plots represents the dynamic of the variable in context to other predictors. There is a distinct difference between the perceptron models (MLP and MLPE) compared to the rest (XGB, RF and TABPFN) regarding binary variables. The firstly mentioned models value all instances of a positive or negative value equally, meaning that for instance having a vegetation on TEE/TTE adds a certain probability of having IE regardless of other variables. For XGB, RF and TABPFN a binary variable has a wider range of impact on the final model. This indicated that these models are utilizing non-linear relationships more considerably than their counterparts, which was one of the major objectives for this thesis.

Hyperparameters

The optimized hyperparameters indicate that a high level of regularization is beneficial for the model, which is expected due to the small sample size. For RF the more aggressive restrictor of maximum features, $\log_2$, is used and XGB preferred subsampling and feature restrictions below 1.0. We can also deduce that the RF model more consistently chooses similar maximum depths, while XGB has more variety, and the perceptron models has a wide range of nodes in the hidden layer for different folds. The stability of the RF model could indicate that it was better fit for tuning of hyperparameters which is reflected in the performance. Lastly, one can interpret that the SVM models in the majority of cases chose RBF over the linear model which indicates that there is some advantage in using a non-linear kernel.

4.2 Conclusion

The ML models outperformed the linear logistic regression model significantly for MLP, MLPE, RF, XGB and TABPFN, but SVM did not reach significance. They were exceeded in sensitivity and specificity by the Duke-ISCVID criteria in the single point. The best performing ML algorithms to diagnose IE was TABPFN followed by RF. This is largely due to the small sample size. The ML models largely utilized known predictors of IE but also identified new important variables including eGFR, CRP, no known focus and BMI. Episodes with evidence of IE on echocardiography where the diagnosis was rejected by the endocarditis team offered difficulties for the models that need to be further explored.

4.3 Future Work

This thesis is among the first of its kind in the context of utilizing ML to diagnose IE, investigate which algorithm is most effective, and to identify new possible predictive factors. The findings we have presented offer new possibilities in ruling in favor or against the diagnosis, but further research is needed to enable their applications.

To enable usage of ML in a clinical setting, future studies need to include larger datasets to fully utilize the power of ML. In our work many of the ML algorithms fall short due instability or overtraining. To investigate the usability of for instance MLP, MLPE and maybe deeper neural networks more episodes of IE must be included. Further research should also aim to include further variables upon which

the endocarditis team bases their decision, such as immunosuppression and more complex descriptions of echocardiography findings.

A possible new way of approaching this problem could be to look at only episodes of suspected IE where there is no vegetation on TEE/TTE. As we have seen this variable is the dominating feature importance and to help decide when treatment is needed when this factor is not present has clinical implications.

Lastly, future work should focus on improving model interpretability and identifying which variables interact to drive predictions. To be able to apply the knowledge acquired the final product must either be a rigidly tested ML algorithm with input that is easy to compile and use by clinicians, or new variables or relationships that can be included in future criteria or another easily understandable diagnostic scoring tool.

4.4 Critical Evaluation

This thesis was largely able to achieve its set objective, but the methodology had several limitations. While being able to outperform the logistic regression model, the ML models were succeeded by the Duke-ISCVID criteria.

As discussed previously in the thesis, a larger sample size would be needed to fully take advantage of the ML algorithms, especially the perceptron models. To gather information from medical records is a time consuming and expensive task which offers challenges in this context.

Another limitation which partly explains the shortcoming of exceeding the Duke criteria is that important variables are not included in the model. As discussed, all information upon which the endocarditis team makes their decision is not included. When designing a model with the aim of being easily interpretable there needs to be a limitation of variables, but not to the extent that it compromises the performance.

The ML models offered new possible predictive variables, but a limitation is the lack of definite statistic test for feature importances. We can find correlation but the estimate of what the probability of making these findings by chance is difficult. On the contrary, appropriate regularization methods were used which should mitigate most random effects in the models. The feature importance also highlights that the models heavily rely on echocardiography which is another limitation.

To find applications for the ML models the future number of variables must be reduced, but as this is an early stage in the development the wide range of input features is a strength. It enables the selection of important features without exposing the risk of excluding important ones.

Another strength is the usage of a multidisciplinary team as target variable. The alternative of using the Duke-ISCVID criteria would introduce a large problem in overfitting to that exact model.

Lastly, the application of machine learning models that are trained and tested on separate data offers a novelty which hopes to inspire more research on the subject.

5. Bibliography

1. Momtazmanesh S, Saeedi Moghaddam S, Malakan Rad E, Azadnajafabad S, Ebrahimi N, Mohammadi E, et al. Global, regional, and national burden and quality of care index of endocarditis: the global burden of disease study 1990-2019. *Eur J Prev Cardiol.* 2022;29(8):1287-97.
2. 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(39):3222-32.
3. Evangelista A, Gonzalez-Alujas MT. Echocardiography in infective endocarditis. *Heart.* 2004;90(6):614-7.
4. Botelho-Nevers E, Thuny F, Casalta JP, Richet H, Gouriet F, Collart F, et al. Dramatic reduction in infective endocarditis-related mortality with a management-based approach. *Arch Intern Med.* 2009;169(14):1290-8.
5. Fowler VG, Durack DT, Selton-Suty C, Athan E, Bayer AS, Chamis AL, et al. The 2023 Duke-International Society for Cardiovascular Infectious Diseases Criteria for Infective Endocarditis: Updating the Modified Duke Criteria. *Clin Infect Dis.* 2023;77(4):518-26.
6. Lindberg H, Berge A, Jovanovic-Stjernqvist M, Hagstrand Aldman M, Krus D, Oberg J, et al. Performance of the 2023 Duke-International Society of Cardiovascular Infectious Diseases Diagnostic Criteria for Infective Endocarditis in Relation to the Modified Duke Criteria and to Clinical Management-Reanalysis of Retrospective Bacteremia Cohorts. *Clin Infect Dis.* 2024;78(4):956-63.
7. Murdoch DR, Corey GR, Hoen B, Miró JM, Fowler VG, Jr, Bayer AS, et al. Clinical Presentation, Etiology, and Outcome of Infective Endocarditis in the 21st Century: The International Collaboration on Endocarditis–Prospective Cohort Study. *Archives of Internal Medicine.* 2009;169(5):463-73.
8. IBM. What is machine learning? 2025 [May 26, 2025]. Available from: <https://www.ibm.com/think/topics/machine-learning>.
9. Ten Hove D, Slart R, Glaudemans A, Postma DF, Gomes A, Swart LE, et al. 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. *Eur J Nucl Med Mol Imaging.* 2024;51(13):3924-33.
10. Lai CK, Leung E, He Y, Ching-Chun C, Oliver MOY, Qinze Y, et al. A Machine Learning-Based Risk Score for Prediction of Infective Endocarditis Among Patients With Staphylococcus aureus Bacteremia-The SABIER Score. *J Infect Dis.* 2024;230(3):606-13.

11. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res.* 2011;12(null):2825–30.
12. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; San Francisco, California, USA: Association for Computing Machinery; 2016. p. 785–94.
13. Hollmann N, Müller S, Eggenberger K, Hutter F. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:220701848.* 2022.
14. Bouckaert RR, Frank E, editors. *Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms.* *Advances in Knowledge Discovery and Data Mining*; 2004 2004//; Berlin, Heidelberg: Springer Berlin Heidelberg.
15. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems.* 2017;30.
16. Xu H, Kinfu KA, LeVine W, Panda S, Dey J, Ainsworth M, et al. When are deep networks really better than decision forests at small sample sizes, and how? *arXiv preprint arXiv:210813637.* 2021.
17. Habib G, Lancellotti P, Antunes MJ, Bongiorni MG, Casalta JP, Del Zotti F, et al. 2015 ESC Guidelines for the management of infective endocarditis: The Task Force for the Management of Infective Endocarditis of the European Society of Cardiology (ESC). Endorsed by: European Association for Cardio-Thoracic Surgery (EACTS), the European Association of Nuclear Medicine (EANM). *Eur Heart J.* 2015;36(44):3075-128.
18. Pyo WK, Kim HJ, Han K, Kim JN, Lee SJ, Kim JH, et al. The risk of infective endocarditis according to blood pressure in patients with diabetes: a nationwide population-based study. *Clin Hypertens.* 2024;30(1):37.

6. Appendix

6.1 Appendix 1

The Duke-ISCVID criteria from 2023 for diagnosing IE. Abbreviations: ISCVID - International Society for Cardiovascular Infectious Diseases. IE – infectious endocarditis. HACEK - haemophilusarter, *Aggregatibacter*, *Cardiobacterium hominis*, *Eikenella corrodens*, *Kingella kingae*.

Pathologic criteria:

Conclusional evidence of IE observed during cardiac surgery or by identification of microorganisms or histologic changes.

Major criteria:

A) Microbiological Evidence

- Positive blood cultures with typical IE microorganisms, such as *Staphylococcus aureus*, viridans-group streptococci, *Streptococcus gallolyticus*, HACEK group, or community-acquired enterococci, from two separate blood culture sets, or
- Three or more separate blood-cultures with atypical pathogens.

B) Imaging Evidence

- Echocardiography or cardiac CT showing vegetation, abscess, pseudoaneurysm, intracardiac fistula, or dehiscence of a prosthetic valve, or
- 18F-FDG PET/CT indicating abnormal metabolic activity involving a native or prosthetic valve (at least 3 months post-implantation), ascending aortic graft (with valve involvement), intracardiac device leads, or other prosthetic material.

C) Surgical Findings

- Evidence of IE observed during cardiac surgery if not major imaging nor pathologic criteria is fulfilled.

Minor criteria

- a) Predisposition: Presence of predisposing conditions, such as prosthetic valve, previous IE, congenital heart disease, intravenous drug use, CIED or more than mild regurgitation or stenosis of any etiology.
- b) Fever: Temperature $\geq 38.0^{\circ}\text{C}$.
- c) Vascular Phenomena: Major arterial emboli, septic pulmonary infarcts, mycotic aneurysms, intracranial hemorrhage, conjunctival hemorrhages, or Janeway lesions.
- d) Immunologic Phenomena: Glomerulonephritis, Osler's nodes, Roth's spots, or rheumatoid factor.
- e) Minor microbiology: Positive blood cultures that do not meet the major criterion or serologic evidence of active infection with an organism consistent with IE.
- f) Immunoimaging: Positive 18F-FDG PET/CT within 3 months of implantation of prosthetic material.

Definite IE: 2 major + ≥ 0 minor, 1 major + ≥ 3 minor, 0 major + ≥ 5 minor, or pathologic criteria.

Possible IE: 1 major + ≥ 1 minor or 0 major + ≥ 3 minor

6.2 Appendix 2

The definition for all variables used.

Variable	Definition
Predisposing factors	
Age / years	Age at discharge.
Sex (female)	As defined in personal security number.
BMI	Body mass index.
Native valve disease	Above mild stenosis or regurgitation before admission.
CIED	Cardiovascular implantable electronic device, including permanent pacemakers, implantable cardioverter defibrillators, and cardiac resynchronization therapy devices.
Previous IE	Previously recognized infective endocarditis.
Intravenous drug use	Active intravenous drug use at admission.
Prosthesis	Valvular prosthesis in any valve.
Bio prosthesis	Prosthesis made of biological material.
Mechanical prosthesis	Prosthesis made of metal.
Baseline eGFR	The best estimated glomerular flow rate the last two years before admission.
Clinical findings	
Vegetation on TEE/TTE	Vegetation on echocardiography.
Abscess on TEE/TTE	Abscess on echocardiography.
Leakage on TEE/TTE	Leakage on echocardiography.
Finding on CT	Finding corresponds with infective endocarditis in cardiac-CT.
Finding on FDG-PET/CT	Finding corresponds with infective endocarditis in FDG-PET/CT.
Heart murmur	Heart murmur on auscultation. Previous known or unknown.
Seven days symptoms	Seven days of symptom agreeing with infective endocarditis before admission.
Emboli	Peripheral or central emboli.
Days of fever	Days of continuous fever during admission. One day of normal temperature allowed.
Maximum CRP	Maximum C-reactive protein during the first five days of admission.
CRP on day five	C-reactive protein on the fifth day of admission.
Kidney injury	Two or of three of hematuria, proteinuria, and urine sediment.
Acute eGFR	Maximum estimated glomerular flow rate during admission.
Infectious focus	
Focus: spondylodiscitis	Infectious focus in disc or vertebrae.
Focus: none	No known infectious focus other than cardiac.
Focus: joint or bone	Infectious focus in joint or bone.
Focus: lung	Infectious focus in lung, including pneumonia and pneumonitis.
Focus: urinary tract	Infectious focus in the urinary tract.
Focus: gastrointestinal	Infectious focus in the gastrointestinal tract.

Focus: teeth	Infectious focus in teeth.
Focus: muscle	Infectious focus in muscle.
Focus: other	Other infectious focus.
Blood cultures	
Number of positive bottles	Number of positive bottles.
Polymicrobial bacteremia	Two or more pathogens found in blood cultures.
Negative culture	No pathogen found in blood culture.
Persistent bacteremia	Two positive blood cultures separated by 48-72 hours with adequate antibiotic therapy.
Acquisition type	
Community acquired	Acquired outside hospital or healthcare environment.
Healthcare associated	The patient visited a hospital or received specialized nursing care within the last 30 days, received acute at hospital treatment for two or more days within the last 90 days, or resided at a nursing home.
Nosocomial	Blood culture was positive 48 hours after admission to hospital.

6.3 Appendix 3

The best hyperparameters for each machine learning model. Maximum features describes how many features are considered when splitting a node as a function of all features (square root or log base 2). Subsample describes the fraction of the data that is randomly used for each boosting round. Colsample by tree tells the fraction of features used for each tree.

<i>Model</i>	<i>Hyperparameter</i>	<i>Best values</i>
<i>LR</i>	C	0.61, 2.64, 1.85, 0.58, 7.66, 2.81, 0.81, 0.49, 0.42, 6.09
	L1 ratio	0.36, 0.78, 0.77, 0.00, 0.46, 0.49, 0.50, 0.028, 0.37, 0.17
<i>SVM</i>	C	1.95, 0.16, 1.34, 1.40, 1.32, 1.01, 1.61, 2.50, 0.20, 0.16
	kernel	RBF, linear, RBF, RBF, RBF, RBF, RBF, RBF, linear, linear
<i>MLP</i>	hidden layer size	3, 10, 3, 3, 7, 3, 7, 10, 2, 2
	alpha	0.48, 0.33, 0.39, 0.52, 0.34, 0.42, 0.29, 0.39, 0.45, 0.31
	batch size	32, 16, 16, 32, 32, 16, 16, 32, 16
	initial learning rate	0.05, 0.05, 0.05, 0.05, 0.02, 0.05, 0.02, 0.05, 0.05, 0.05
	solver	All SGD
<i>MLPE</i>	hidden layer size	2, 5, 2, 7, 7, 3, 10, 7, 7, 5
	alpha	0.24, 0.44, 0.40, 0.32, 0.48, 0.20, 0.37, 0.39, 0.45, 0.36
	batch size	16, 16, 16, 16, 32, 16, 16, 32, 16, 16
	initial learning rate	0.05, 0.02, 0.02, 0.01, 0.05, 0.05, 0.02, 0.02, 0.02, 0.01
	solver	All SGD
<i>RF</i>	maximum depth	8, 9, 7, 8, 8, 8, 9, 9, 9, 7
	minimum samples split	2, 4, 2, 4, 3, 2, 4, 3, 3, 3
	number of estimators	200, 2000, 2000, 500, 500, 200, 500, 2000, 200, 100
	maximum features	log2, log2, sqrt, log2, sqrt, log2, log2, sqrt, log2, sqrt
<i>XGB</i>	maximum depth	4, 6, 2, 3, 3, 6, 3, 3, 8, 8
	minimum child weight	2, 1, 5, 1, 1, 10, 5, 1, 2, 1
	subsample	1.0, 0.7, 0.7, 0.5, 0.7, 1.0, 1.0, 0.5, 0.7, 1.0
	colsample by tree	0.25, 0.25, 0.75, 0.5, 0.5, 0.5, 0.25, 0.5, 0.25, 0.5
	learning rate	0.05, 0.05, 0.05, 0.005, 0.05, 0.005, 0.05, 0.005, 0.05, 0.01
	number of estimators	100, 500, 50, 300, 100, 500, 100, 300, 100, 500

6.4 Appendix 4

Median area under curve for the receiver operating characteristic curve (ROC-AUC) for MLP and RF when trained with Duke-ISCVID classification as target variable. The testing was done with 5 folds repeated twice (10 samples). Interquartile range is shown in parenthesis.

Model	ROC-AUC for Duke-ISCVID
MLP	0.96 (0.95-0.96)
RF	0.96 (0.95-0.96)

Lund University Department of Automatic Control Box 118 SE-221 00 Lund Sweden		<i>Document name</i> BACHELO'S THESIS	
		<i>Date of issue</i> June 2025	
		<i>Document Number</i> TFRT-7672	
<i>Author(s)</i> Axel Svenningsson		<i>Supervisor</i> Mattias Ohlsson, Centre for Environmental and Climate Science (CEC), Lund University, Sweden Magnus Rasmussen, Department of Clinical Sciences, Lund University, Sweden Bo Bernhardsson, Dept. of Automatic Control, Lund University, Sweden (examiner)	
<i>Title and subtitle</i> Novel Applications of Machine Learning to Guide Diagnosis of Endocarditis			
<i>Abstract</i> Infective endocarditis (IE) is a severe disease affecting the inner lining of the heart. Its diverse symptoms complicate diagnosis and the current main diagnostic tool is the Duke criteria which classifies suspected cases as definite, possible or rejected IE. A substantial portion of cases falls into the ambiguous "possible" category and there is a need for a more precise diagnostic system. This thesis aimed to investigate the usage of machine learning (ML) models, utilizing non-linear relationships, as a tool to determine whether a patient suffers from IE and extract what features are important in decision making. Multiple ML algorithm techniques were trained with a set of variables extracted from medical records. For target variable the opinion from a team consisting of medical specialists of multiple disciplines was used. The results showed that all ML algorithms outperformed logistic regression, which was used as a linear model for comparison. The Duke criteria in turn outperformed the ML models in specificity and sensitivity. The two best performing models were based on tabular prior-data fitted network and random forest. The ML models largely utilized known predictors of IE but also identified new important variables including estimated glomerular flow rate, C-reactive peptide, no known focus and body mass index.			
<i>Keywords</i>			
<i>Classification system and/or index terms (if any)</i>			
<i>Supplementary bibliographical information</i>			
<i>ISSN and key title</i> 0280-5316		<i>ISBN</i>	
<i>Language</i> English	<i>Number of pages</i> 1-46	<i>Recipient's notes</i>	
<i>Security classification</i>			

<http://www.control.lth.se/publications/>