Barreñada *et al.*
Diagnostic and Prognostic Research (2024) 8:14
https://doi.org/10.1186/s41512-024-00177-1

Diagnostic and Prognostic Research

RESEARCH

Open Access

Understanding overfitting in random forest for probability estimation: a visualization and simulation study

Lasai Barreñada[1,2], Paula Dhiman[3], Dirk Timmerman[1,4], Anne-Laure Boulesteix[5] and Ben Van Calster[1,2,6*]

Abstract

Background Random forests have become popular for clinical risk prediction modeling. In a case study on predicting ovarian malignancy, we observed training AUCs close to 1. Although this suggests overfitting, performance was competitive on test data. We aimed to understand the behavior of random forests for probability estimation by (1) visualizing data space in three real-world case studies and (2) a simulation study.

Methods For the case studies, multinomial risk estimates were visualized using heatmaps in a 2-dimensional subspace. The simulation study included 48 logistic data-generating mechanisms (DGM), varying the predictor distribution, the number of predictors, the correlation between predictors, the true AUC, and the strength of true predictors. For each DGM, 1000 training datasets of size 200 or 4000 with binary outcomes were simulated, and random forest models were trained with minimum node size 2 or 20 using the ranger R package, resulting in 192 scenarios in total. Model performance was evaluated on large test datasets ($N = 100{,}000$).

Results The visualizations suggested that the model learned "spikes of probability" around events in the training set. A cluster of events created a bigger peak or plateau (signal), isolated events local peaks (noise). In the simulation study, median training AUCs were between 0.97 and 1 unless there were 4 binary predictors or 16 binary predictors with a minimum node size of 20. The median discrimination loss, i.e., the difference between the median test AUC and the true AUC, was 0.025 (range 0.00 to 0.13). Median training AUCs had Spearman correlations of around 0.70 with discrimination loss. Median test AUCs were higher with higher events per variable, higher minimum node size, and binary predictors. Median training calibration slopes were always above 1 and were not correlated with median test slopes across scenarios (Spearman correlation − 0.11). Median test slopes were higher with higher true AUC, higher minimum node size, and higher sample size.

Conclusions Random forests learn local probability peaks that often yield near perfect training AUCs without strongly affecting AUCs on test data. When the aim is probability estimation, the simulation results go against the common recommendation to use fully grown trees in random forest models.

Keywords Random Forest, Prediction modeling, Risk estimation

*Correspondence:
Ben Van Calster
ben.vancalster@kuleuven.be
Full list of author information is available at the end of the article

Background

Random Forests (RF) is an ensemble learning method introduced by Leo Breiman in 2001 [1]. The difference between RF and other tree ensemble methods such as bagging or boosting is that the trees in RF are independent. A bootstrap sample is selected at each tree and at each node of each tree a random subset of predictors is considered for the best split. This reduces the correlation between trees.

RF is used across a variety of clinical problems [2] and in recent years it has become very popular for clinical prediction modeling [3–6]. Its popularity has risen due to its good reported performance in applied studies, and its claimed robustness against overfitting in combination with the limited need for hyperparameter tuning [7]. It has been reported that RF models have better performance when the individual trees in the ensemble are overfitted [8–10]. Although RF has been widely investigated as a "classifier", the literature about their performance as probability estimation trees (PET) is scarce.

In a recent study of women with an ovarian tumor, we compared the performance of different machine learning algorithms to estimate the probability of five tumor types (benign, borderline malignant, stage I primary invasive, stage II–IV primary invasive, and secondary metastatic) [11]. We developed prediction models on training data ($n = 5909$) using multinomial logistic regression (MLR), ridge multinomial logistic regression, RF, XGBoost, neural networks, and support vector machines with Gaussian kernel. We evaluated discrimination performance using the Polytomous Discrimination Index (PDI) as a multiclass area under the receiver operating characteristic curve (AUC) [12, 13]. The PDI is the probability that, when presented with a random patient from each category, the model can correctly identify the patient from a randomly selected category. With five outcome categories, the PDI is 1/5 for an uninformative or random model and 1 for a model with perfect discrimination. We observed that RF had near-perfect discrimination on the training data (PDI 0.93 for RF vs 0.47–0.70 for other models), and was competitive on the external validation data (PDI 0.54 for RF vs 0.41–0.55 for other models; $n = 3199$) (Table S1). The observation that the RF model had near-perfect (i.e., highly suspicious) discrimination on training data, yet performed competitively during external validation, may be somewhat counterintuitive. Such high training set performance suggests strong overfitting by modeling considerable amounts of noise, which would lead to reduced performance on new data [14, 15]. This was an interesting observation for us: the training results suggest a suspiciously high degree of overfitting by RF compared to other models, such that we would have expected a stronger reduction in performance on new data for RF. In this study, we aimed to understand the behavior of random forests for probability estimation by (1) visualizing data space in three real world case studies and (2) conducting a simulation study.

The paper outline is as follows. In the "Random forest for probability estimation" section, we summarize the RF algorithm for probability estimation, in the "Case studies" section, we visualize the predictions for the ovarian tumor data, and present two additional case studies. In the "Simulation study" section, we present a simulation study to explore the effect of tree depth and training sample size and the data generation mechanism (DGM) to better understand the behavior of the RF algorithm. In the "Overall discussion" section, we discuss our findings.

Random forest for probability estimation

When the outcome is categorical, RF can be used for classification or probability estimation. In this work we will use random forest for probability estimation [16, 17]. RF is a tree-based ensemble method, and when used for probability estimation it works as follows:

1. Draw *ntree* bootstrap samples from the original training dataset, where *ntree* denotes the number of trees in the forest.
2. On each bootstrap sample, construct a tree using recursive binary splits. To reduce the correlation between trees, a number of predictors (*mtry*) are chosen randomly at each split. *mtry* is a hyperparameter and can be tuned, but often a default value equal to the square root of the total predictors (P) is used. A split on one of these predictors is chosen so that the selected splitting criterion (e.g., Gini index) is optimized.
3. Splits are consecutively created as long as all child nodes contain a specific minimum number of observations (*min.node.size*). When a node cannot be split without violating this condition, the node becomes a final leaf node. Other stopping criteria can be defined [18]. For each leaf node, the proportion of cases from each outcome class can be calculated. Alternatively, the majority vote can be determined: the outcome class that has the most cases in the leaf.
4. To obtain a probability estimate for each outcome class i for a new case, we first determine the new case's appropriate leaf node for each of the *ntree* trees. Then, two basic approaches are possible. The first uses the proportion of the *ntree* majority votes (cf step 3) that equal i. The second averages the proportion of cases from class i (cf step 3) across the *ntree* trees [16].

In the seminal books "The Elements of Statistical Learning" and "An Introduction to Statistical Learning", the authors highlight the simplicity of training RF models [14, 19]. Regarding the commonly encountered claim that RF cannot overfit, the authors indicate that increasing *ntree* does not cause overfitting. It has been suggested that *ntree* does not need to be tuned, but that too low values lead to suboptimal performance [7, 20]. A value of 500 or even 250 has shown to be sufficient in most applications [7]. A typical value for *mtry* is $\sqrt{P}$, as recommended by Breiman, or lower values to maximize decorrelation [14]. Hastie and colleagues suggest that *min.node.size* can be set to a very low value, even 1 and that *mtry* is a more important tuning parameter: "when the number of variables is large, but the fraction of relevant variables small, random forests are likely to perform poorly with small mtry. … Our experience is that using full-grown trees seldom costs much, and results in one less tuning parameter" [14].

Case studies

Methods

We aimed to visualize the estimated probabilities in data space to obtain a better understanding of the phenomenon where RF models with near-perfect discrimination also performed competitively during external validation. We followed a typical random train-test split used in machine learning procedures. We developed RF and MLR prediction models on the training set using two continuous and a number of categorical predictors. We use only two continuous predictors because if we set the categorical predictors to a fixed value, e.g., the most common one, we can show a two-dimensional subset of the complete data space by showing the two continuous predictors on the x-axis and y-axis. We can show estimated probabilities in this subset as a heatmap, and show individual cases (from training or test set) as a scatter plot. This allows us to visualize how RF and MLR transform predictor values into probability estimates, for example in terms of smoothness. Obviously, only cases for which the categorical values equal the chosen fixed value can be shown. By choosing different fixed values for categorical variables, we can visualize different subsets of data space. We noticed that the range of estimated probabilities was larger for RF than MLR. Therefore, to ensure a proper visualization of the high- and low-risk estimates, the greyscale in the heatmaps is bounded to the minimum and maximum predicted probabilities by each model in each panel. We also include figures using the same scale for all heatmaps in Additional file 1.

The RF models were trained with ranger package, with $ntree = 500$, $mtry = \lceil \sqrt{P} \rceil$, and min.node.size = 2. Ranger estimates the probabilities with Malley's probability machine methods which averages the proportion of cases from each class over the terminal nodes from each of the trees [16, 21]. In MLR models, we modeled continuous predictors using restricted cubic splines (rcs) with 3 knots to allow nonlinear associations [22, 23]. For each model, we calculated the train and test PDI and multinomial calibration plots. The code for training the models and generating the plots is available in the OSF repository (https://osf.io/y5tqv/).

Ovarian cancer diagnosis

This prospective study collected data on patients between 1999 and 2012. All patients had at least one adnexal (ovarian, para-ovarian, or tubal) mass that was judged not to be a physiological cyst, provided consent for transvaginal ultrasound examination, were not pregnant, and underwent surgical removal of the adnexal mass within 120 days after the ultrasound examination. We randomly split the data ($N = 8398$) into training ($n = 5900$, 70%) and test parts ($n = 2498$, 30%), and developed models on the training data using patient age (in years) and CA125 (in IU/L) as continuous variables, and five ultrasound based categorical variables (proportion of solid tissue, number of papillary projections, if the mass has more than 10 locules, if the mass has shadows and if the mass has ascites). Note that the proportion of solid tissue is a continuous variable that can be seen as a semi-categorical variable with 75% of observations having values 0 or 1. The distribution of classes in the dataset was 66% (5524) for benign tumors, 6% (531) for a borderline ovarian tumor, 6% (529) for stage I ovarian cancer, 17% (1434) for stage II–IV ovarian cancer, and 5% (380) for metastatic cancer to the ovaries (detailed information in Table S2). The apparent PDI was 0.97 for RF and 0.52 for MLR. In the test set the PDI decreased to 0.56 for the RF model and remained 0.52 for the MLR.

Figure 1 shows heatmaps for the estimated probabilities of a benign, borderline, stage I invasive, stage II–IV invasive, and secondary metastatic tumor, with training data cases superimposed (see Figures S1–2 for extended visualizations). Cases belonging to the class to which the probabilities refer are shown in red, and other cases in green. One set of heatmaps refers to the fitted RF model, and the other to the fitted MLR model. Whereas estimated probabilities from the regression model change smoothly according to the values of the continuous predictors, the estimated probabilities from the RF model peak where events from the training data were located. Where many events were found in proximity, these peaks combined into a larger area with increased probability. For events in less densely

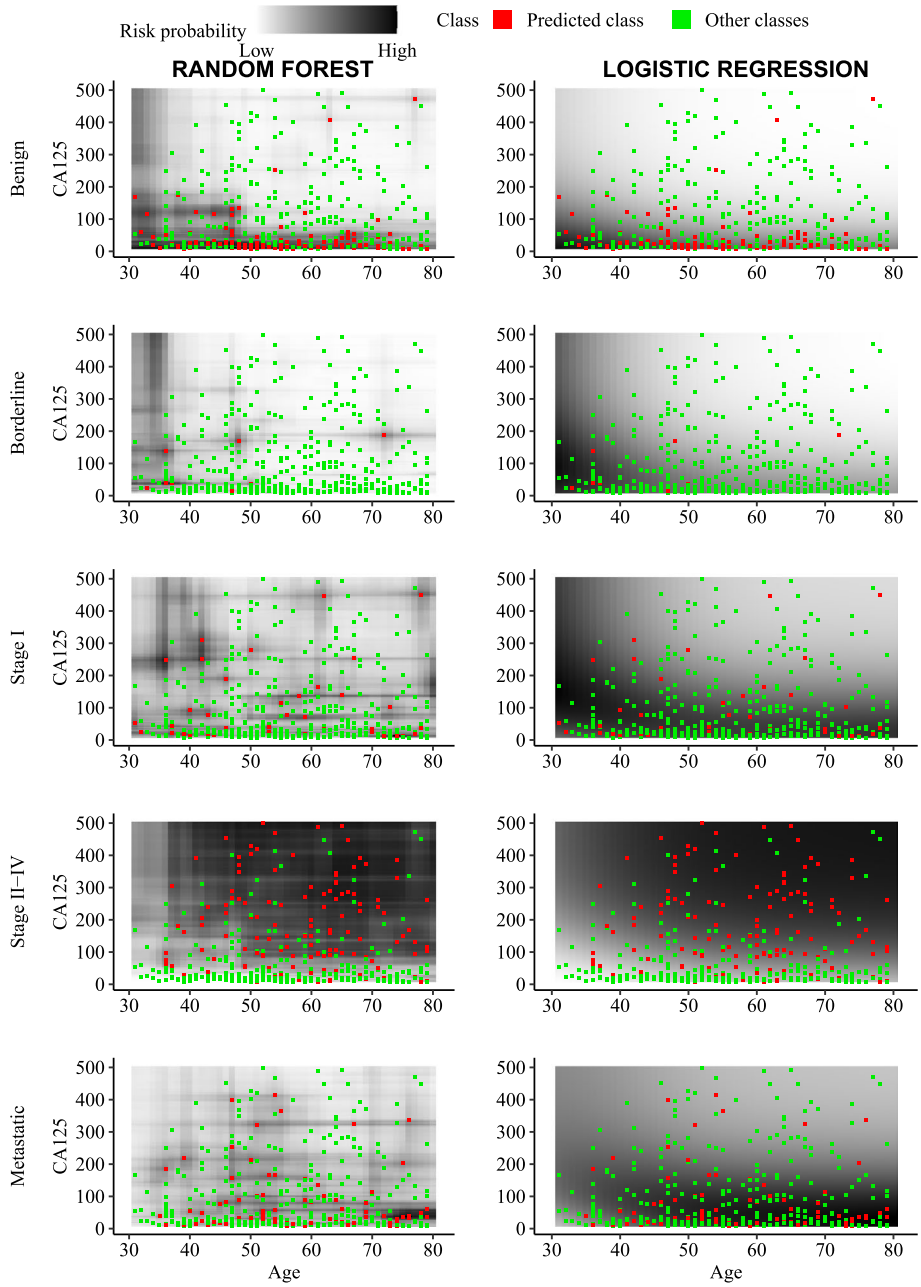

Fig. 1 Random Forest and logistic regression probability estimation in data space for 4 subtypes of ovarian malignancy diagnosis with cases in training set superimposed. CA125 bounded to 500

populated areas of data space, these peaks were idiosyncratic: in test data, events in these areas of data space tend to be located in different places (Fig. 2 and Figures S3–S4). The calibration performance of the RF model was very poor in the training set: high probabilities were underestimated and low probabilities were overestimated (Fig. 3). Calibration in the test set was much better.

CRASH3: traumatic brain injury prognosis

CRASH-3 data was collected between 2012 and 2019 for a multicenter, randomized, placebo-controlled trial to measure the effects of tranexamic acid on death, disability, vascular occlusive events, and other morbidities in 12,660 patients with acute traumatic brain injury (TBI) [24]. We used age (years) and systolic blood pressure (mmHg) as continuous variables and sex, Glasgow

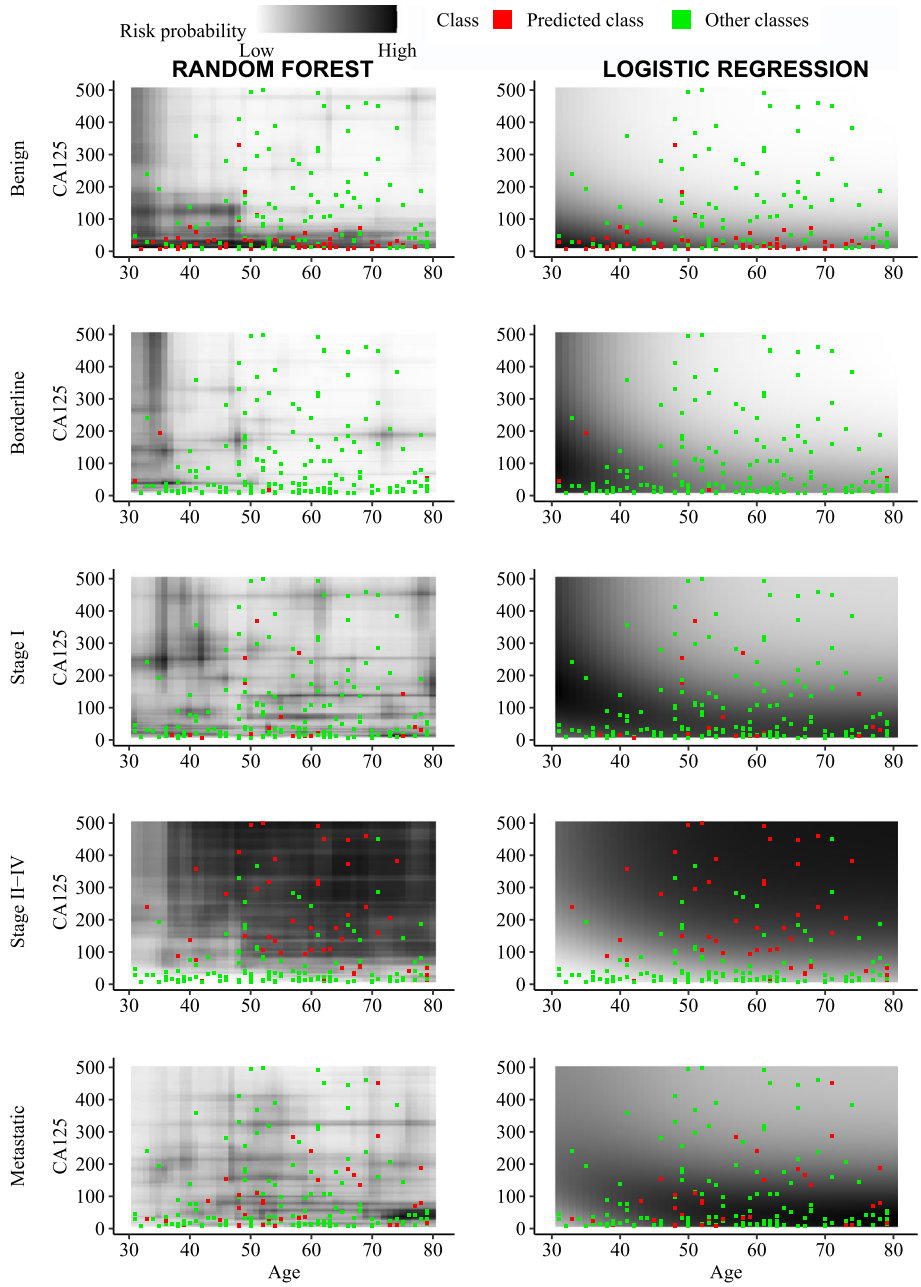

Fig. 2 Random forest and logistic regression probability estimation in data space for 4 subtypes of ovarian malignancy diagnosis with cases in test set superimposed. CA125 bounded to 500

Coma Scale (GCS) eye-opening (4 levels), and pupillary reaction (4 levels) as categorical variables. We performed a complete case analysis (CCA) removing patients for which one or more values were missing obtaining a complete dataset of 12,548 patients. CCA was used for simplicity and because the phenomenon under study should not be affected importantly by this. The outcome was measured 28 days after randomization: alive ($n=10{,}022$, 80%), death due to head injury ($n=2309$, 18%), or death of other cause ($n=217$, 2%) (detailed information in Table S3). The training set included 8783 patients (70%), and the test set was 3765 (30%).

For RF, the PDI was 0.96 in train data and 0.54 in test data. For MLR, the PDI was 0.61 and 0.60, respectively.

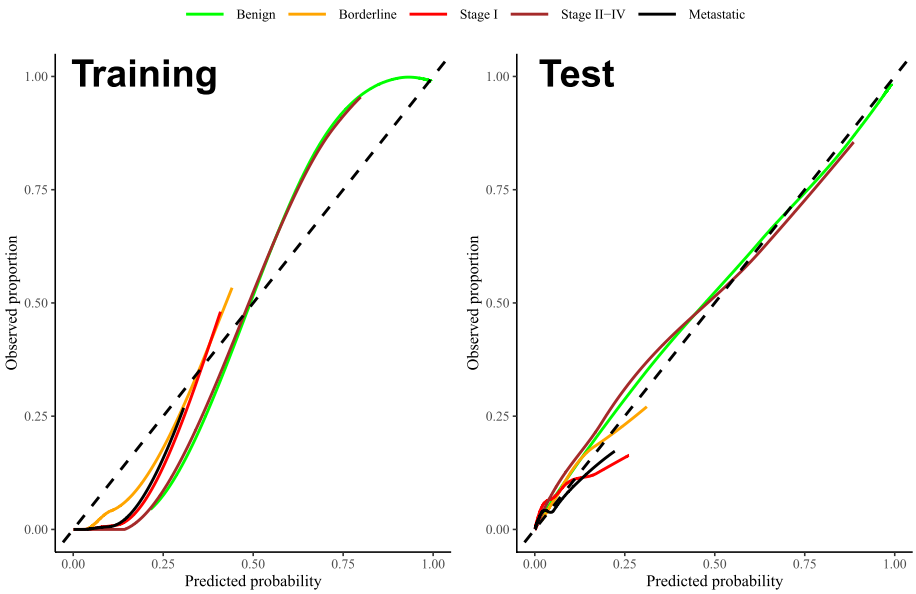

Fig. 3 Calibration plots for random forest model in ovarian cancer data. Observed proportion is estimated with a LOESS model. The plots only show observed proportions for predicted probabilities between quantiles 5th and 95th

The heatmaps drew a similar picture compared with the previous case study: RF had clear probability peaks whilst MLR had smoothly changing probabilities (Figures S5–S8). The calibration plots for RF were also similar: poor calibration in training data but decent in the test data for the 2 most common outcomes (Figure S9).

IST: type of stroke diagnosis

The International Stroke Trial (IST) database was designed with the aim of establishing whether early administration of aspirin, heparin, or both or neither influenced the clinical course of acute ischaemic stroke [25]. Data for the 19,435 patients with suspected acute ischaemic stroke were recruited between 1992 and 1996. We use age (years) and systolic blood pressure (mmHg) as continuous variables, and conscious state (fully alert vs drowsy), deficit of face (yes vs no), deficit of arm/hand (yes vs no), deficit of leg/foot (yes vs no), dysphasia (yes vs no), and hemianopia (yes vs no) as categorical variables. We again performed a CCA retaining 15,141 patients. The outcome is the type of stroke: ischaemic ($n=13{,}622$, 90%), indeterminate ($n=736$, 5%), hemorrhagic ($n=439$, 3%), or no stroke ($n=344$, 2%) (detailed information in Table S4). The training set included 10,598 patients (70%), and the test set 4543 (30%).

The RF model had a training PDI of 0.89 and a test PDI of 0.35. For MLR, the training set PDI was 0.39, the test set PDI was 0.41. In this dataspace, the phenomenon is notorious, with very local peaks around train cases (FiguresS10–S13). The calibration plots for training and test data showed poor calibration (Figure S14).

Simulation study

Aim

We conducted a simulation study to assess which key factors of the modeling setup (dataset and minimum node size) contribute to the phenomenon of having an exaggerated AUC in the training data without strong signs of overfitting in test data. We report the simulation study using the ADEMP (aims, data-generating mechanisms, estimands, methods, and performance measures) structure [26]. The code for the simulation study can be found in the OSF repository (https://osf.io/y5tqv/).

Data-generating mechanism (DGM)

For the simulation study, we generated data by assuming that the true model was in the form of an MLR with an outcome event fraction of 0.2 (see Additional file 1: Appendix 1: Simulation Algorithm for details).

The 48 DGMs differed according to the following parameters:

i. *Predictor distribution*: predictors were either all continuous with multivariate normal distribution or all binary with 50% prevalence.
ii. *Number of predictors*: there were either 4 true predictors (0 noise predictors), 16 true predictors (0 noise predictors), or 16 predictors of which

12 noise predictors. Noise predictors had a true regression coefficient of 0.

iii. *Correlation between predictors*: Pearson correlations between all predictors were either 0 or 0.4.

iv. *True AUC*: this was either 0.75 or 0.90.

v. *Balance of regression coefficients*: the true model coefficients that were not 0 were either all the same (balanced) or not (imbalanced). When imbalanced, one-fourth of the predictors have a coefficient that is 4 times larger than the others.

The true model coefficients for generating the data were obtained by trial error and are available in Table S5.

Estimands

For both training and test data, we estimate model discrimination, whether risk estimates have too high (overconfidence) or too low (underconfidence) spread and the prediction error. Underconfidence reflects the situation in Fig. 3 (Train): high probabilities are underestimated, and low probabilities are overestimated. Overconfidence is the opposite: high probabilities are overestimated, and low probabilities are underestimated. For test data, we also calculate discrimination loss vs the true model and the relative contribution of bias and variance to the prediction error. As the training and test samples are based on the same DGM, the test results reflect internal rather than external validation.

Methods

We fitted models on training datasets of size 200 (small) or 4000 (large), and for RF models we used values for *min.node.size* of 2 or 20 and ranger package for training. As a result, there were $2\times2\times3\times2\times2\times2\times2=192$ scenarios. For each scenario, 1000 simulation runs were performed (i.e., 1000 different training datasets). For RF, *ntree* was fixed at 500, and *mtry* at the square root of the number of predictors (default value). Models were validated on a single large test dataset per DGM ($N=100{,}000$) to avoid sampling variability.

Performance

For discrimination we calculated the AUC and for confidence of risk estimates the calibration slope (slope < 1 means overconfidence, slope > 1 underconfidence). The calibration slope is calculated as the slope of a logistic regression (LR) model fitting the outcome to the logit of the estimated probabilities as the only predictor. Calibration intercept is calculated fitting the same model for a slope of 1 by setting the predicted probabilities as an offset term. Discrimination loss was calculated as the difference between true AUC and median test AUC. Finally, the mean squared error (MSE) of the predicted probabilities was calculated according to [27] as the sum of squared bias and variance (see Additional file 1: Appendix 2: Simulation metrics for details).

Results

The aggregated simulation results using median and interquartile range for discrimination and calibration and mean and standard deviation for mean squared error are available in Additional file 1 (Table S6). The complete simulation including the code and 1000 simulations for each of the 192 scenarios is available in the OSF repository (https://osf.io/y5tqv/).

Discrimination

In the simulation study, the median training AUCs were close to 1 in most of the cases. The median training AUC was between 0.97 and 1 unless there were 4 binary predictors, or 16 binary predictors combined with a minimum node size of 20 (Fig. 4). Higher *min.node.size* resulted in less extreme training AUCs.

In general, median test AUCs were higher when there was a large vs small training dataset, high vs low *min.node.size*, high vs low correlation between predictors, binary versus continuous predictors, and 4 versus 16 predictors (except with correlated continuous predictors) (Fig. 5). All other simulation factors being equal, the scenarios with 4 true and 12 noise predictors had results for the AUC that was identical to scenarios with 16 predictors (Figs. 4 and 5 and S15–16).

The Spearman correlation between median training AUC and discrimination loss was 0.72 for scenarios with a true AUC of 0.9, and 0.69 for scenarios with a true AUC of 0.75 (Figure S17). The median discrimination loss was 0.025 (range 0.00 to 0.13). In the 114 scenarios where the median training AUC was ≥ 0.99, the median discrimination loss was 0.036 (range 0.003 to 0.13). In the other scenarios, the median discrimination loss was 0.013 (0.00 to 0.069).

Calibration

Median training calibration slopes ranged between 1.10 and 19.4 (Fig. 6 and Figure S18): the probability estimates were always underconfident where high probabilities were underestimated and low probabilities overestimated. This is the consequence of perfect separation between events and non-events in training data (i.e., AUC = 1) which means that any estimation above 0 or below 1 is underconfident (Figure S19). The median slope was lowest in scenarios with few binary predictors or higher min.node.size. Median test calibration slopes ranged between 0.45 and 2.34. Across all scenarios, the Spearman correlation between the median training

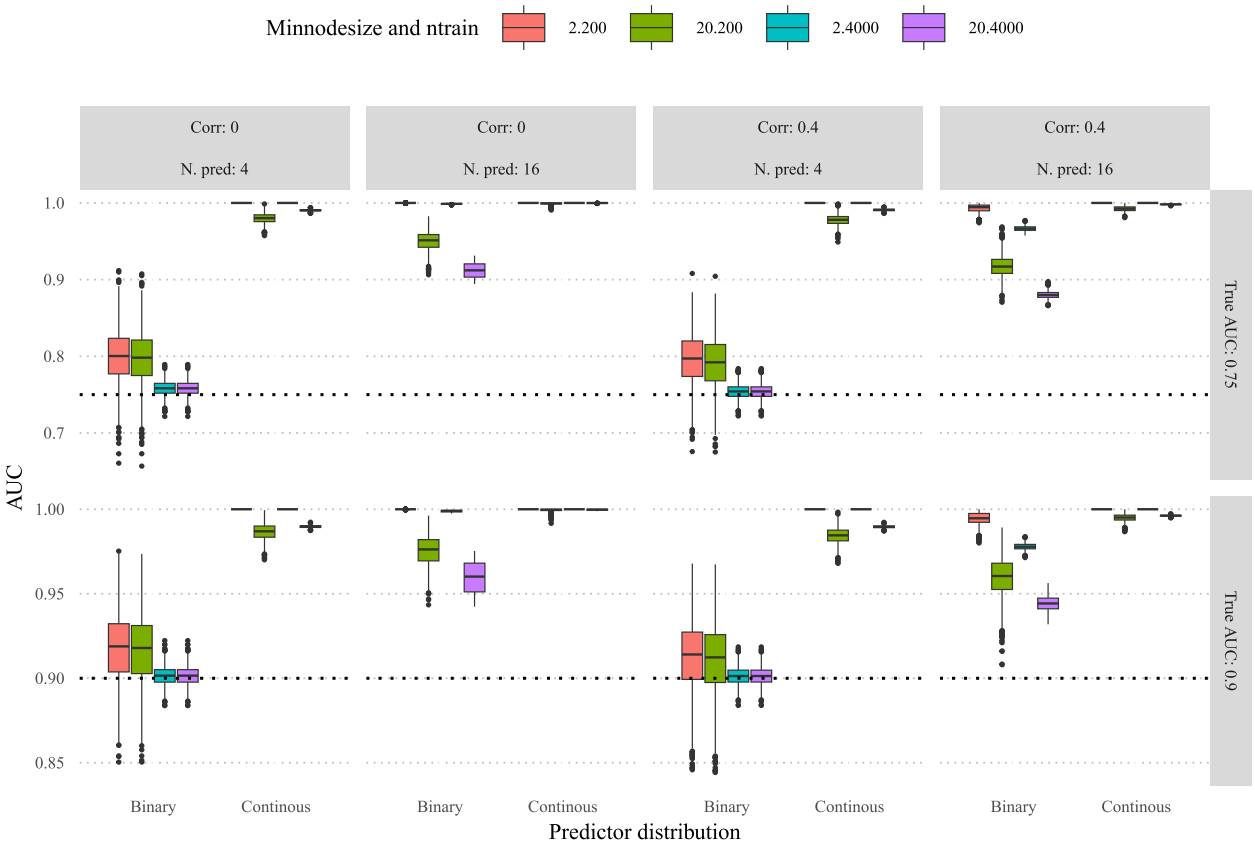

Fig. 4 Training AUC by simulation factors and modeling hyperparameters in scenarios without noise. Scenarios are aggregated by strength because this simulation factor had minimal effect

slope and median test slope was −0.11 (Figure S17). Median test slopes were mainly higher when the true AUC or min.node.size was higher. In addition, median test slopes tended to be higher with binary predictors, uncorrelated predictors, and higher sample sizes (Fig. 7 and Figure S18). Calibration slopes were similar in scenarios with 16 true predictors, 4 true predictors, and 12 noise predictors (Figs. 6 and 7 and Figure S20–21). In the 78 scenarios without perfect training (AUC < 0.99) the median test calibration slope was between 0.59 and 2.34 with a median of 1.10. In the 114 scenarios with almost perfect training AUC (≥ 0.99), the median calibration slope was 0.92.

Mean squared error (MSE)

Median MSE across scenarios was 0.008 (range 0.000–0.045) with a median squared bias of 0.002 (range 0.000–0.038) and median variance of 0.005 (range 0.000–0.017). For the 114 scenarios with median training AUC ≥ 0.99, we observed a median test MSE of 0.010 with a median squared bias of 0.004 (range 0.0004–0.0384) and a median variance of 0.006 (range 0.001–0.017). For the rest of the scenarios, the median test MSE was 0.006. Across all scenarios, the Spearman correlation of mean test squared bias and mean test variance with median training AUC were 0.47 and 0.43, respectively. The correlation with the discrimination loss was 0.51 for the squared bias and 0.70 for the variance. Lower sample size in training was associated with higher median test variance, and with higher median test squared bias in scenarios with continuous predictors. Lower min.node.size (i.e., deeper trees) was associated with a lower variance but higher bias in test data when the training sample size was small. More predictors were associated with higher bias, whereas the correlation between predictors and higher true AUC was associated with lower bias (Fig. 8). The models with noise predictors had lower variance and higher bias compared to scenarios with 4 true and no noise predictors (Figure S22).

Overall discussion

We tried to better understand and visualize the behavior of random forests for probability estimation. We make three key observations from this work. First, RF models learn local probability peaks around training set events, in particular when the trees are very deep (i.e., low *min.*

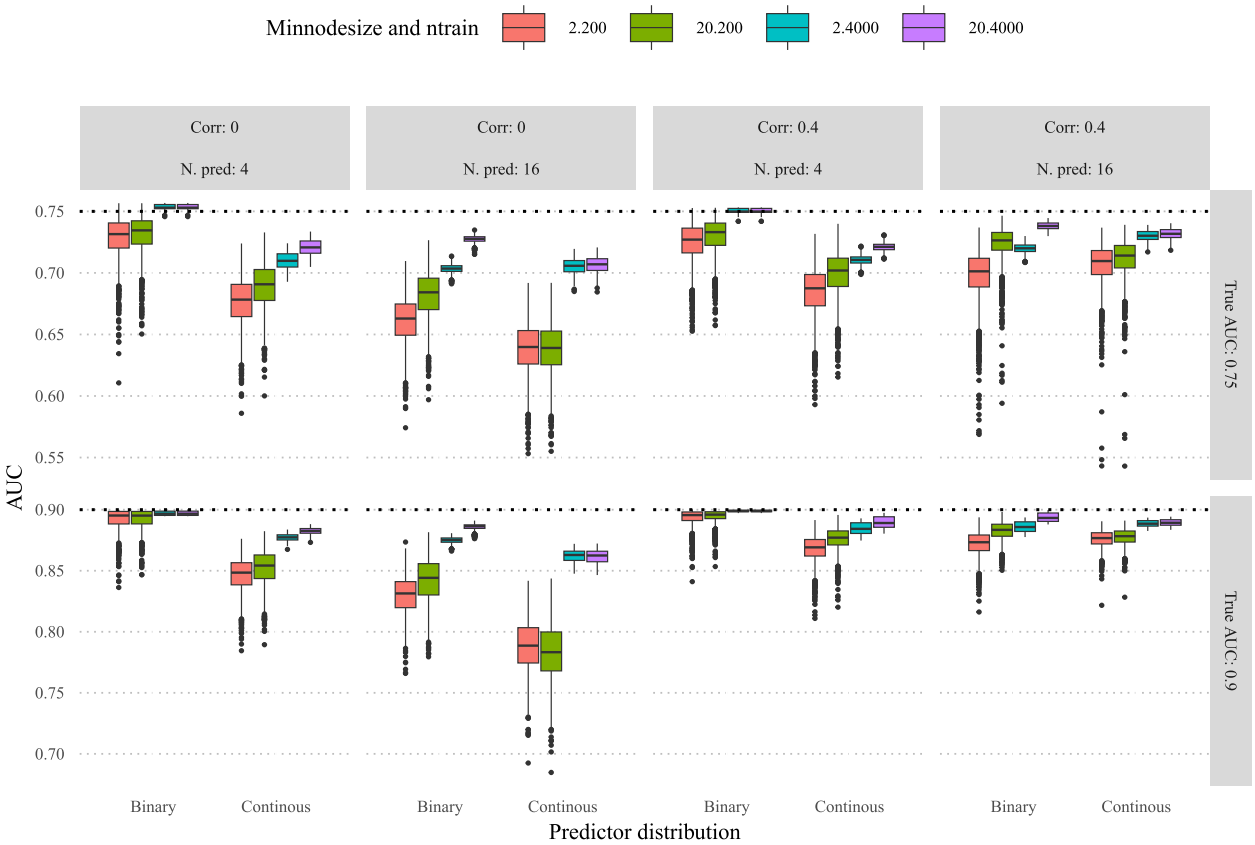

Fig. 5 Test AUC by simulation factors and modeling hyperparameters in scenarios without noise. Scenarios are aggregated by strength

node.size) and when there are continuous predictors. Where a group of events is located close to one another in 'data space', the probability peaks are combined into a region of increased probability. Where events are isolated in data space, the probability peaks are very local. Learning through peaks leads to very optimistic (often near perfect) discrimination in training data, but also to reduced discrimination in new data compared to models that rely on less deep trees (cf. simulation settings where RF models with *min.node.size* 2 vs 20 were fitted on a set of continuous predictors). Probably, the reduction in discrimination on new data is modest because the local peaks for isolated events are often harmless for new data: it is unlikely to see an event in the exact same location. Second, RF also suffers from "classical overfitting" in which models with higher discrimination on training data tend to have lower discrimination on new data than models with less optimistic discrimination (cf. simulation settings where RF models are learned on 4 binary predictors using 200 vs 4000 training cases). Third, in training and test data, calibration performance for RF models is different from what we commonly observe for LR models. Whereas LR based on maximum likelihood leads by definition to calibration slopes of 1 on training data, calibration slopes for RF models were always above 1 on training data. Also, as opposed to LR, calibration slopes for RF models do not converge to 1 on new data. This different behavior is probably caused by the pragmatic way in which probabilities are obtained for RF models, whilst LR estimates probabilities in a principled way through maximum likelihood.

The simulation results regarding discrimination and calibration go against fitting very deep trees when using RF for probability estimation. This is in line with recent work that illustrated that RF using deeply grown trees results in risk estimates that are particularly unstable [28]. The heatmaps for our case studies illustrate how RF models with deeply grown trees lead to probabilities that change non-smoothly with changes in the values of predictors. It has been suggested to set *min.node.size* to 5% or 10% of the sample size [16, 29]. Alternatively, *min.node.size* can be tuned. Although this was not the focus of our study, it seems natural to tune with logloss (also known as the negative loglikelihood or cross-entropy) or Brier score as the loss function since they capture calibration [18]. In their work, Ledger and colleagues tuned *min.node.size* by optimizing logloss based on tenfold cross-validation for 30 random values for mtry and *min.*

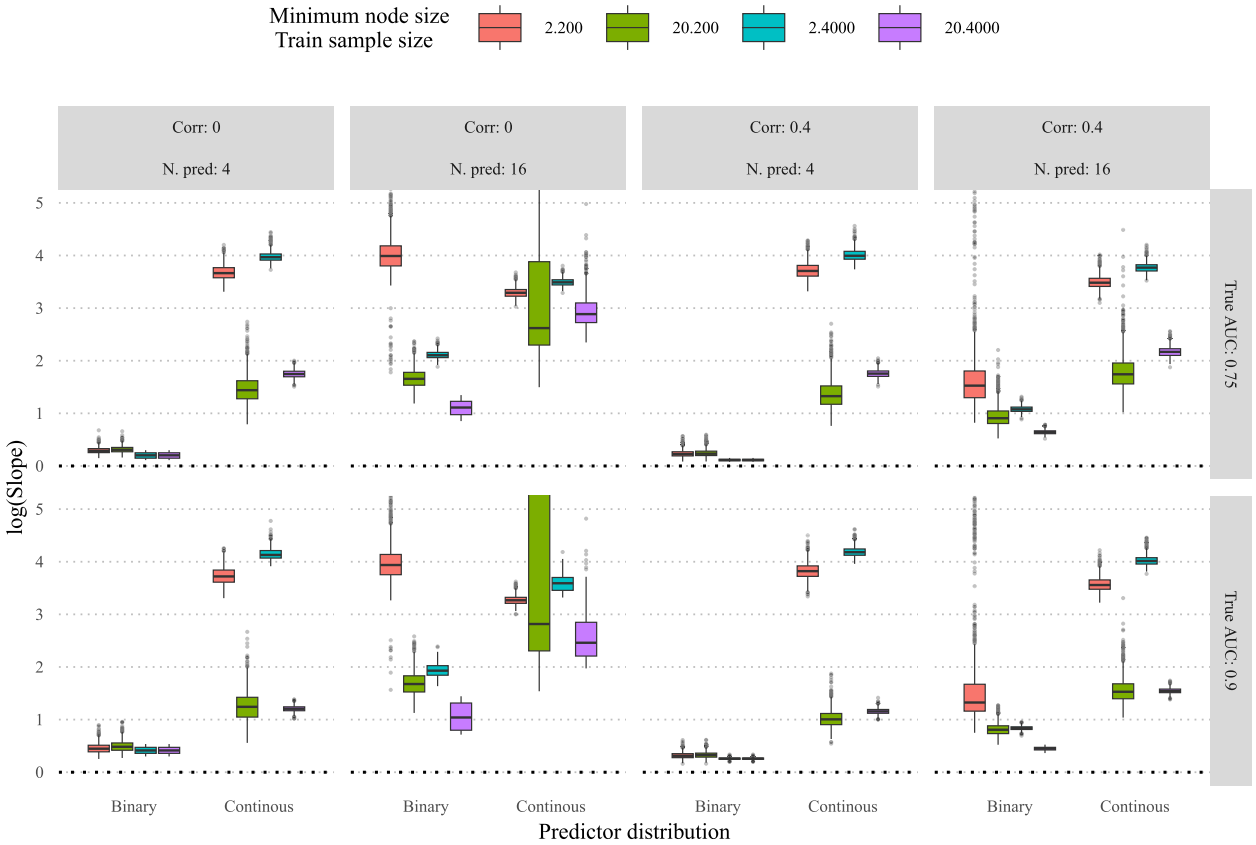

Fig. 6 Training set calibration log slope by simulation factors and modeling hyperparameters in scenarios without noise. Scenarios are aggregated by strength. The ideal value for the log slope is 0

node.size using the trainControl and train functions from the caret R package [11]. This resulted in *mtry* = 3 and *min.node.size* = 15, with competitive results in test data. Applying the tuneRanger R package with 200 iterations for our case studies based on logloss yielded an optimal *min.node.size* of 8 (0.1% of training set size) for ovarian cancer data, 261 (2.1%) for CRASH3 data and 591 (3.9%) for IST data [18]. These values are higher than the default values in many statistical software programs.

Three comments regarding the interpretation of discrimination and calibration results for RF models are worth making. First, it is well known that apparent performance, i.e., performance assessment on the exact same dataset that was used to train the model, is overly optimistic [30, 31]. Our work indicates that this is a fortiori case for RF. Unfortunately, some studies present their RF models with "excellent" performance because they only present discrimination for the training data [4, 32, 33]. Instead, proper internal and external validation results should be reported. Second, from a regression modeling perspective, a calibration slope that is clearly below 1 on internal validation is a symptom of overfitting. This cannot be applied in the same way for RF models. Models based on larger training samples had higher calibration slopes in our simulation study, but a calibration slope of 1 does not appear to have a special meaning. Models with a calibration slope above 1 on test data when trained on 200 training samples, had an even higher slope when trained on 4000 samples. The interpretation of risk estimates based on RF requires caution, probably because probabilities are generated in a very ad hoc way. Of course, the calibration slope still quantifies in a descriptive manner whether the risk estimates are on average too confident (slope < 1), not confident enough (slope > 1), or fine (slope = 1). Third, despite that RF models had high discrimination results in the training data (suggestive of overfitting), the calibration slope in the training data was always above 1 (risk estimates show too little spread, suggestive of underfitting). This appears to be a consequence of the bootstrapping procedure in combination with the low *min.node.size*. Due to the bootstrapping, a training set case was part of approximately 63% of bootstrap samples and therefore was used for 63% of the trees in the forest. When averaging the proportion of events in the appropriate leaf nodes over

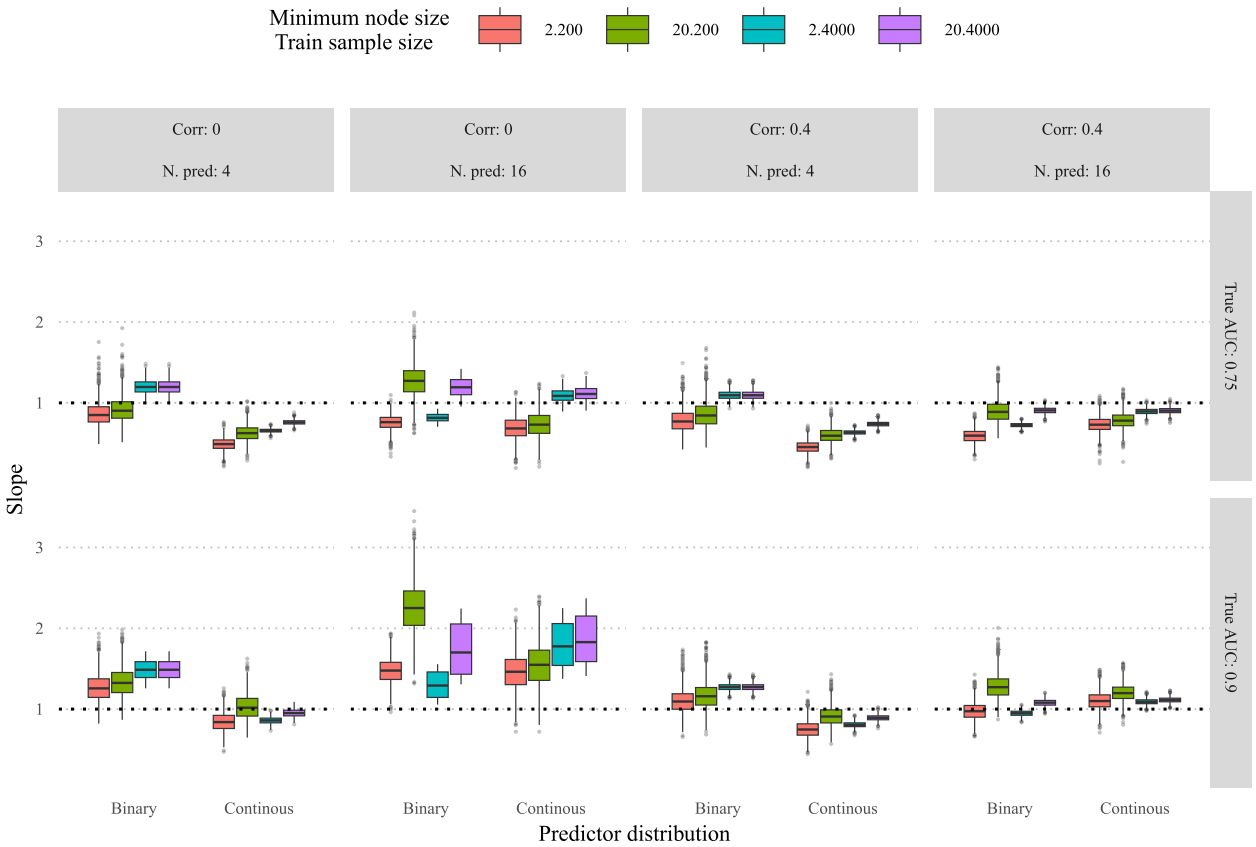

Fig. 7 Test set calibration slope by simulation factors and modeling hyperparameters in scenarios without noise. Scenarios are aggregated by strength. The ideal value for the slope is 1

all *ntree* trees to get a probability estimate for a given training set case, 63% of these proportions are near perfect: close to 1 if the training case is an event, close to 0 if the training cases is a non-event. The remaining 37% are more variable and often far less good. The 63% near-perfect proportions cause discrimination to be very high because most events will end up with an estimated probability of an event that is higher than that for most non-events. The remaining 37% of the proportions pull the probability estimates away from 0 (if the case is a non-event) or 1 (if the case is an event), leading to calibration slopes > 1.

Although the aim of our paper was largely educational, it links to previous more fundamental work and fills a gap in the literature by explicitly studying factors that contribute to better discrimination and calibration of new data. Wyner and colleagues (2017) argued that RF has excellent performance because it is an "interpolating classifier", i.e., it is fitted with little to no error to the train data [34]. They argue that the interpolation should not be confused with overfitting. Even if the individual trees are overfitting, each training set case is not used in about 37% of the individual trees, such that averaging over trees partially solves overfitting. Belkin and colleagues have linked this to a double descent curve for highly flexible algorithms: when the complexity of the model increases, test set performance first improves, then deteriorates, and finally improves again once the 'interpolation threshold' (where perfect training performance is achieved) is exceeded [35]. Recently, however, Buschjäeger and Morik opposed the existence of double descent in RF [36]. Mentch and Zhou linked the success of RF to the signal-to-noise ratio (SNR) of the data. In their work, they present that the randomness of RF is beneficial in situations with low SNR whereas bagging is preferred if SNR is high [37]. They explain the success of RF by the low SNR of many real-world datasets. This view contradicts the view that flexible algorithms work best when the SNR is high [38]. Finally, the issue of calibration of estimated probabilities in the context of RF models received little attention in the literature, although it is key for optimal clinical decision making [39]. It has been suggested that using fully grown trees in RF leads to suboptimal risk estimates [16, 29]. However, this is rarely mentioned and hence it is common to see that low minimum node sizes are recommended because the problem is treated as a classification

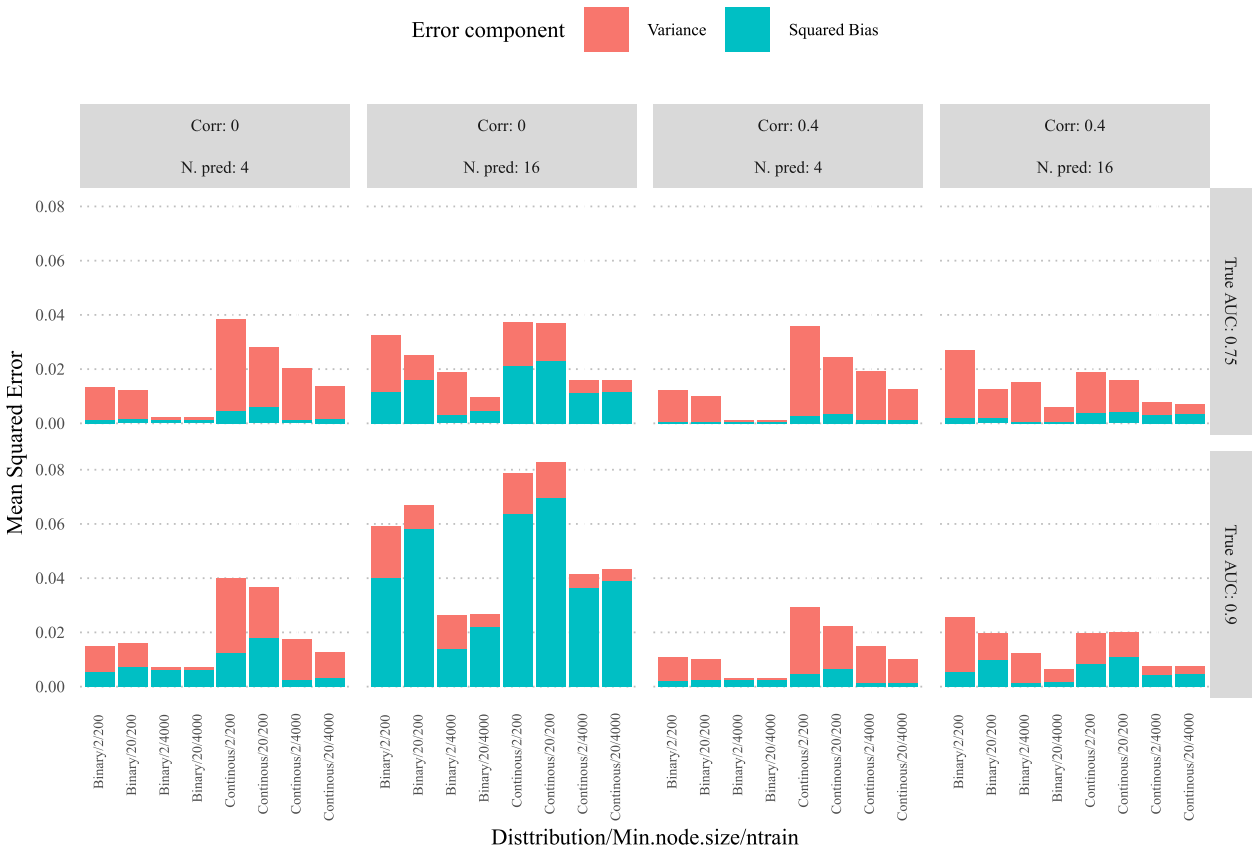

Fig. 8 Mean squared error across scenarios without noise aggregated by strength

problem instead of a probability estimation problem (e.g., see default for randomForest package or scikit-learn). We think that the current study sheds further light on probability estimation in RF. Of course, a generic alternative that works for any miscalibrated model is to recalibrate the probabilities of the RF afterward using new data [40].

We identified the following limitations of our study. Firstly, a simulation study is always limited by the included scenarios. It would be of interest to include more simulation factors and values per simulation factor (e.g., for *min.node.size*) in the simulation study or to include scenarios where RF hyperparameters are tuned rather than fixed. Tuning could improve the calibration of the models [18, 41]. However, the simulation study already had 192 scenarios, and adding more factors or values would increase the computational cost exponentially and would overcomplicate the interpretation of the results. Topics that could be investigated in further simulation studies include varying other hyperparameters than *min.node.soze* (e.g., *mtry*, sampling fraction, splitting rule) and investigating more values for sample size, number of predictors, the proportion of noise predictors, and min.node.size. Secondly, we were using logistic DGMs without nonlinear or nonadditive associations with the outcome. We assumed that the impact of this would be limited, because the focus was not on the comparison of RF with LR, and any nonlinearity or nonadditivity (including the absence of it) has to be learned by the algorithm. Thirdly, the traditional RF algorithm selects variables at each split in a way that favors continuous over binary variables [42]. Continuous variables can split in many ways, such that there is often a split that, perhaps by chance, has a better Gini impurity reduction than the Gini impurity reduction for a binary variable. The splits for continuous variables may often overfit, thereby increasing training discrimination but decreasing test discrimination. It is well documented that this affects variable importance measures [42], but it may also be relevant for model performance. The problem can be addressed by using an adapted RF algorithm such as cforest from the partykit package [43]. These adapted algorithms grow conditional inference trees (CIT) instead of classification trees. For the case studies, we observed that tuning and using the adapted RF yields a less optimistic training AUC and similar or slightly better test performance (see OSF Repository, https://osf.io/y5tqv/).

However, our aim was to understand how the characteristics of the data and the modeling process affected the models, hence we did not systematically explore the effects of tuning or alternative algorithms.

We conclude that RF tends to exhibit local overfitting by learning probability peaks, in particular when the RF model is based on deeply grown trees. This local overfitting can lead to highly optimistic (near perfect) discrimination on the training data but to reduced discrimination on new data compared to RF models based on less deeply grown trees. In line with the work of Kruppa and colleagues [41], our results go against the recommendation to use fully grown trees when using RF for probability estimation.

Abbreviations

RF	Random forest
PET	Probability estimation tree
MLR	Multinomial logistic regression
PDI	Polytomous Discrimination Index
AUC	Area under the receiver operating characteristic curve
TBI	Traumatic brain injury
GSC	Glasgow coma score
CCA	Complete case analysis
IST	International stroke trial
ADEMP	Aims, data-generating mechanisms, estimands, methods, and performance measures
DGM	Data-generating mechanism
LR	Logistic regression
MSE	Mean squared error
SNR	Signal-to-noise ratio

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s41512-024-00177-1.

Additional file 1: Train and test performance of different machine learning algorithms on predicting ovarian type of tumour in terms of polytomous discrimination index (PDI) and ovarian malignancy in terms of AUC (Benign vs Malignant). AUC, Area Under receiver operating curve; MLR, Multiple Linear Regression; RF, Random Forest; XGBoost, Extreme gradient boosting; NN, Neural network; SVM, Support Vector Machine. Table S2. Distribution of different classes in the ovarian cancer dataset. Table S3. Distribution of different classes in the ovarian CRASH dataset. Table S4. Distribution of different classes in the ovarian IST dataset. Table S5. Coefficients and simulation factors. Table S6. Main simulation results. Figure S1. Random forest probability estimation in data space for ovarian malignancy diagnosis with random forest (left) and multinomial logistic regression (right). Squares refer to train cases. Figure S2. Random forest probability estimation in data space for ovarian malignancy diagnosis with random forest (left) and multinomial logistic regression (right) using same scale for all panels. Squares refer to train cases. Figure S3. Random forest probability estimation in data space for ovarian malignancy diagnosis with random forest (left) and multinomial logistic regression (right). Squares refer to test cases. Figure S4. Random forest probability estimation in data space for ovarian malignancy diagnosis with random forest (left) and multinomial logistic regression (right) using same scale for all panels. Squares refer to test cases. Figure S5 Figure S5 Figure S5. Figure S5. Random forest probability estimation in data space for 3 possible outcomes in CRASH 3 dataset. Squares refer to train cases. Figure S6 Figure S6 Figure S6. Figure S6. Random forest probability estimation in data space for 3 possible outcomes in CRASH 3 dataset using same scale for all panels. Squares refer to train cases. Figure S7. Random forest probability estimation in data space for 3 possible outcomes in CRASH 3 dataset. Squares refer to test cases. Figure S8. Random forest probability estimation in data space for 3 possible outcomes in CRASH 3 dataset. Squares refer to test cases using same scale for all panels. Figure S9. Multinomial calibration plots of CRASH training and test data. Observed proration is estimated with a LOESS model. The plot shows only predicted probabilities between quantiles 5th and 95th. Figure S10. Random forest probability estimation in data space for 4 possible type of strokes in IST dataset. Squares refer to training cases. Figure S11. Random forest probability estimation in data space for 4 possible type of strokes in IST dataset. Squares refer to training cases using same scale for all panels. Figure S12. Random forest probability estimation in data space for 4 possible type of strokes in IST dataset. Squares refer to test cases. Figure S13. Random forest probability estimation in data space for 4 possible type of strokes in IST dataset using same scale for all panels. Squares refer to test cases. Figure S14. Calibration plot in training for IST dataset. Observed proration is estimated with a LOESS model. The plot shows only predicted probabilities between quantiles 5th and 95th. Figure S15. Training AUC by simulation factors and modelling hyperparameters in scenarios with noise. Scenarios are aggregated by strength. Figure S16. Test AUC by simulation factors and modelling hyperparameters in scenarios with noise. Scenarios are aggregated by strength. Figure S17. Spearman correlations of principal metrics across all scenarios, scenarios with true AUCs 0.9 and scenarios with true AUC 0.75. Figure S18. Train and test calibration log(slope) in scenarios without noise. Scenarios are summarised by simulation factors that had minor effect. Figure S19. Histogram of predicted probabilities in training for different simulation scenarios. Minimum node size was always 2 and training sample size 4000. Figure S20. Training set calibration log slope by simulation factors and modelling hyperparameters in scenarios with noise. Scenarios are aggregated by strength. Perfect calibration is 0. Figure S21. Test set calibration slope by simulation factors and modelling hyperparameters in scenarios with noise. Scenarios are aggregated by strength. Perfect calibration is 1. Figure S22. Mean squared error across scenarios with noise aggregated by strength.

Acknowledgements
Partial findings from this paper were presented at the Young Statisticians Meeting 2023 in Leicester (July) and in the Conference of the International Society of Clinical Biostatistics (ISCB) 2023 in Milan (August).

Authors' contributions
Contributions were based on the CRediT taxonomy. Conceptualization: LB, BVC. Funding acquisition: BVC, DT. Project administration: LB. Supervision: BVC. Methodology: LB, PD, ALB, BVC. Resources: LB, DT, BVC. Investigation: LB, PD, DT, ALB, BVC. Validation: LB, BVC. Data curation: LB. Software: LB, BVC. Formal analysis: LB, BVC. Visualization: LB, BVC. Writing—original draft: LB, BVC. Writing—review and editing: all authors. All authors have read, share final responsibility for the decision to submit for publication, and agree to be accountable for all aspects of the work.

Funding
This research was supported by the Research Foundation–Flanders (FWO) under grant G097322N with BVC and DT as supervisors.

Availability of data and materials
All data and code that support the findings of this study are openly available in the OSF repository at (https://osf.io/y5tqv/) [44] except the ovarian cancer dataset which is not publicly available.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Development and Regeneration, Leuven, KU, Belgium. [2]Leuven Unit for Health Technology Assessment Research (LUHTAR), Leuven, KU, Belgium. [3]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK. [4]Department of Obstetrics and Gynecology, University Hospitals Leuven, Leuven, Belgium. [5]Biometry in Molecular Medicine, LMU, Munich, Germany. [6]Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, the Netherlands.

Received: 16 January 2024 Accepted: 17 September 2024
Published online: 27 September 2024

References

1. Breiman L. Random Forests. Mach Learn. 2001;45(1):5–32.
2. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? J Mach Learn Res. 2014;15(90):3133–81.
3. Corradi JP, Thompson S, Mather JF, Waszynski CM, Dicks RS. Prediction of Incident Delirium Using a Random Forest classifier. J Med Syst. 2018;42(12):261.
4. Dai B, Chen RC, Zhu SZ, Zhang WW. Using Random Forest Algorithm for Breast Cancer Diagnosis. In: 2018 International Symposium on Computer, Consumer and Control (IS3C). 2018. p. 449–52.
5. Xu W, Zhang J, Zhang Q, Wei X. Risk prediction of type II diabetes based on random forest model. In: 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB). 2017. p. 382–6.
6. Yao D, Yang J, Zhan X. A Novel Method for Disease Prediction: Hybrid of Random Forest and Multivariate Adaptive Regression Splines. JCP. 2013;8(1):170–7.
7. Oshiro TM, Perez PS, Baranauskas JA. How Many Trees in a Random Forest? In: Perner P, editor. Machine Learning and Data Mining in Pattern Recognition. MLDM 2012. Lecture Notes in Computer Science, vol 7376. Springer, Berlin, Heidelberg; 2012. https://doi.org/10.1007/978-3-642-31537-4_13.
8. Biau G, Scornet E. A random forest guided tour. TEST. 2016;25(2):197–227.
9. Denil M, Matheson D, Freitas ND. Narrowing the Gap: Random Forests In Theory and In Practice. In: Proceedings of the 31st International Conference on Machine Learning. PMLR; 2014 [cited 2023 Aug 14]. p. 665–73. Available from: https://proceedings.mlr.press/v32/denil14.html
10. Breiman L. Some Infinity Theory for Predictor Ensembles | Department of Statistics. J Comb Theory Ser A. 2002;98:175–91.
11. Ledger A, Ceusters J, Valentin L, Testa A, Van Holsbeke C, Franchi D, et al. Multiclass risk models for ovarian malignancy: an illustration of prediction uncertainty due to the choice of algorithm. BMC Med Res Methodol. 2023;23(1):276.
12. Van Calster B, Van Belle V, Vergouwe Y, Timmerman D, Van Huffel S, Steyerberg EW. Extending the c-statistic to nominal polytomous outcomes: the Polytomous Discrimination Index. Stat Med. 2012;31(23):2610–26.
13. Dover DC, Islam S, Westerhout CM, Moore LE, Kaul P, Savu A. Computing the polytomous discrimination index. Stat Med. 2021;40(16):3667–81.
14. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. New York, NY: Springer New York; 2009 [cited 2023 Feb 15]. (Springer Series in Statistics). Available from: http://link.springer.com/https://doi.org/10.1007/978-0-387-84858-7
15. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Cham: Springer International Publishing; 2019 [cited 2023 Feb 14]. (Statistics for Biology and Health). Available from: http://link.springer.com/https://doi.org/10.1007/978-3-030-16399-0
16. Malley JD, Kruppa J, Dasgupta A, Malley KG, Ziegler A. Probability Machines: Consistent Probability Estimation Using Nonparametric Learning Machines. Methods Inf Med. 2012;51(1):74–81.
17. Dankowski T, Ziegler A. Calibrating random forests for probability estimation. Statist Med. 2016;35(22):3949–60.
18. Probst P, Wright MN, Boulesteix AL. Hyperparameters and tuning strategies for random forest. WIREs Data Min Knowl Discovery. 2019;9(3):e1301.
19. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: with Applications in R. New York, NY: Springer US; 2021 [cited 2023 Oct 11]. (Springer Texts in Statistics). Available from: https://link.springer.com/https://doi.org/10.1007/978-1-0716-1418-1
20. Probst P, Boulesteix AL. To Tune or Not to Tune the Number of Trees in Random Forest. J Mach Learn Res. 2018;18(181):1–18.
21. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Softw. 2017;31(77):1–17.
22. Gauthier J, Wu QV, Gooley TA. Cubic splines to model relationships between continuous variables and outcomes: a guide for clinicians. Bone Marrow Transplant. 2020;55(4):675–80.
23. Harrell , FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis [Internet]. Cham: Springer International Publishing; 2015 [cited 2023 Feb 14]. (Springer Series in Statistics). Available from: https://link.springer.com/https://doi.org/10.1007/978-3-319-19425-7
24. CRASH-3 Trial Collaborators. Effects of tranexamic acid on death, disability, vascular occlusive events and other morbidities in patients with acute traumatic brain injury (CRASH-3): a randomised, placebo-controlled trial. Lancet. 2019;394(10210):1713–23.
25. Sandercock PA, Niewada M, Członkowska A. the International Stroke Trial Collaborative Group. The International Stroke Trial database Trials. 2011;12(1):101.
26. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074–102.
27. Friedman JH. On Bias, Variance, 0/1—Loss, and the Curse-of-Dimensionality. Data Min Knowl Disc. 1997;1(1):55–77.
28. Riley RD, Collins GS. Stability of clinical prediction models developed using statistical or machine learning methods. Biometrical Journal. 2023;n/a(n/a):2200302.
29. Kruppa J, Schwarz A, Arminger G, Ziegler A. Consumer credit risk: Individual probability estimates using machine learning. Expert Syst Appl. 2013Oct 1;40(13):5125–31.
30. Altman DG, Royston P. What do we mean by validating a prognostic model? Statist Med. 2000;19(4):453–73.
31. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. Ann Intern Med. 2019;170(1):W1.
32. Chen R, Deng Z, Song Z. The Prediction of Malignant Middle Cerebral Artery Infarction: A Predicting Approach Using Random Forest. J Stroke Cerebrovasc Dis. 2015;24(5):958–64.
33. Yuan H, Fan XS, Jin Y, He JX, Gui Y, Song LY, et al. Development of heart failure risk prediction models based on a multi-marker approach using random forest algorithms. Chin Med J. 2019;132(7):819.
34. Wyner AJ, Olson M, Bleich J, Mease D. Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers. J Mach Learn Res. 2017;18(48):1–33.
35. Belkin M. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. Acta Numer. 2021;30:203–48.
36. Buschjäger S, Morik K. There is no Double-Descent in Random Forests [Internet]. arXiv; 2021 [cited 2023 Jul 25]. Available from: http://arxiv.org/abs/2111.04409
37. Mentch L, Zhou S. Randomization as Regularization: A Degrees of Freedom Explanation for Random Forest Success. J Mach Learn Res. 2020;21(171):1–36.
38. Van Calster B, Wynants L. Machine Learning in Medicine. N Engl J Med. 2019;380(26):2588–90.
39. Van Calster B, Vickers AJ. Calibration of Risk Prediction Models: Impact on Decision-Analytic Performance. Med Decis Making. 2015;35(2):162–9.
40. Ojeda FM, Jansen ML, Thiéry A, Blankenberg S, Weimar C, Schmid M, et al. Calibrating machine learning approaches for probability estimation: A comprehensive comparison. Stat Med. 2023;42(29):5451–78.
41. Kruppa J, Liu Y, Biau G, Kohler M, König IR, Malley JD, et al. Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory. Biom J. 2014;56(4):534–63.
42. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics. 2007;8(1):25.
43. Hothorn T, Hornik K, Zeileis A. Unbiased Recursive Partitioning: A Conditional Inference Framework. J Comput Graph Stat. 2006;15(3):651–74.
44. Barreñada L, Dhiman P, Boulesteix AL, Calster B van. Understanding overfitting in random forest for probability estimation: a visualization and simulation study. 2023 Nov 20 [cited 2024 Jun 28]; Available from: https://osf.io/y5tqv/

Publisher's Note