# Hyperparameter optimization of random forest for effective prediction of water distribution failures

**Vera Bueler-Faudree[1,⋆] and Alan R. Vazquez[2]**

[1]University of California, Los Angeles, and [2]University of Arkansas

⋆ verab@g.ucla.edu

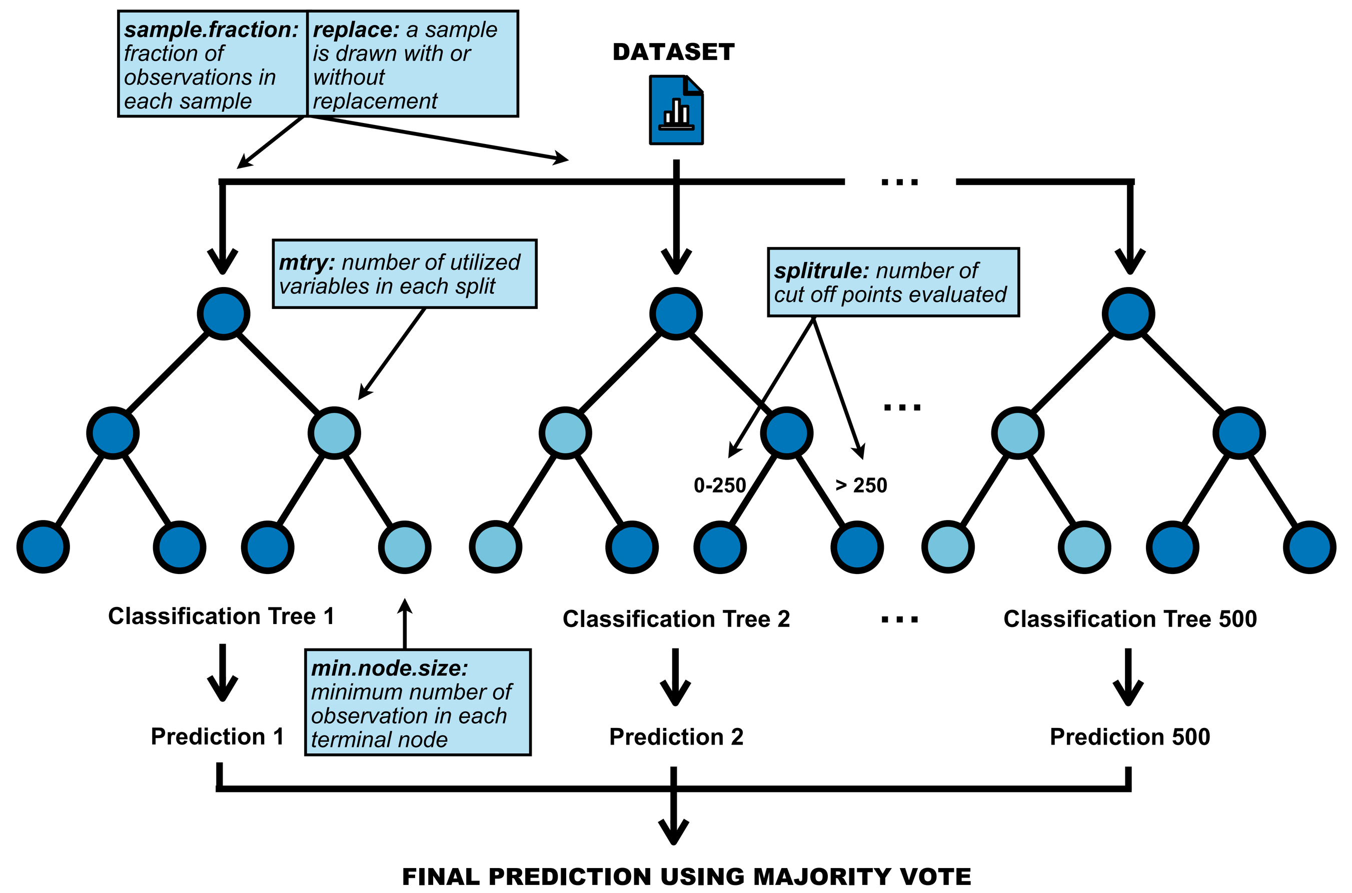
## A WATER DISTRIBUTION SYSTEM IN KENYA

- Western Kenya has a water distribution system of rural pumps that provide clean water to citizens [6].
- Broken pumps prevent entire towns from accessing water.
- Training data with records of 3,934 functioning and 271 non-functioning pumps are available. It has 5 predictors such as number of pumping events, water flow through the pump, and others.

**Goal:** Build a machine learning algorithm to predict pump failures.

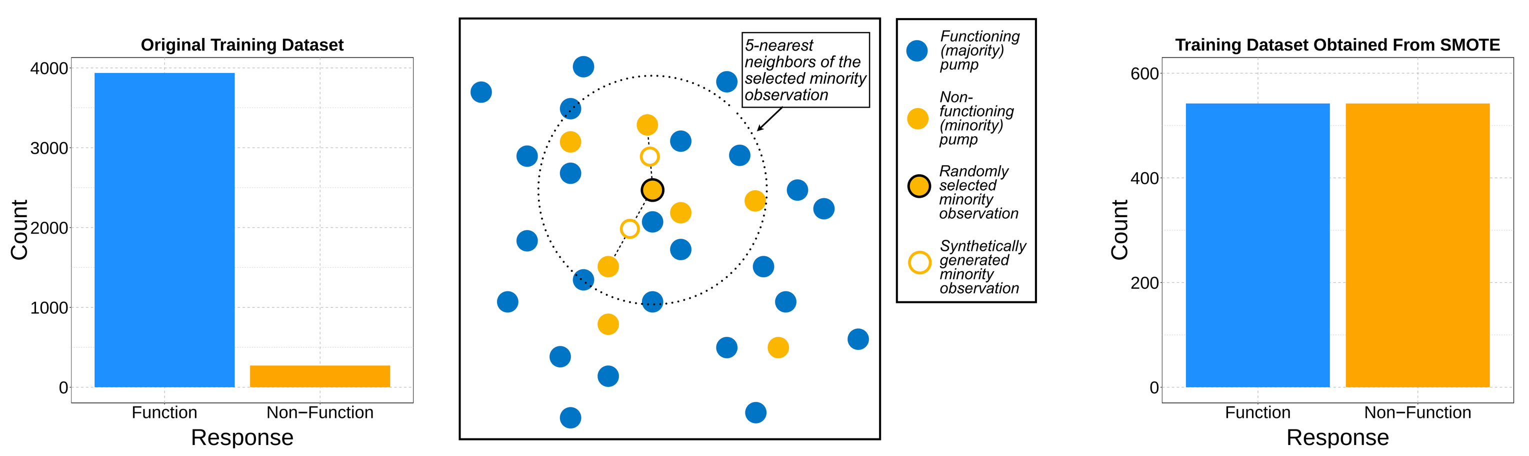
## RANDOM FOREST

Random forest is an ensemble of many classification trees built using random samples of observations and predictors. It has 5 hyperparameters that drive its predictive performance [3, 4].




## SMOTE FOR UNBALANCED DATASETS

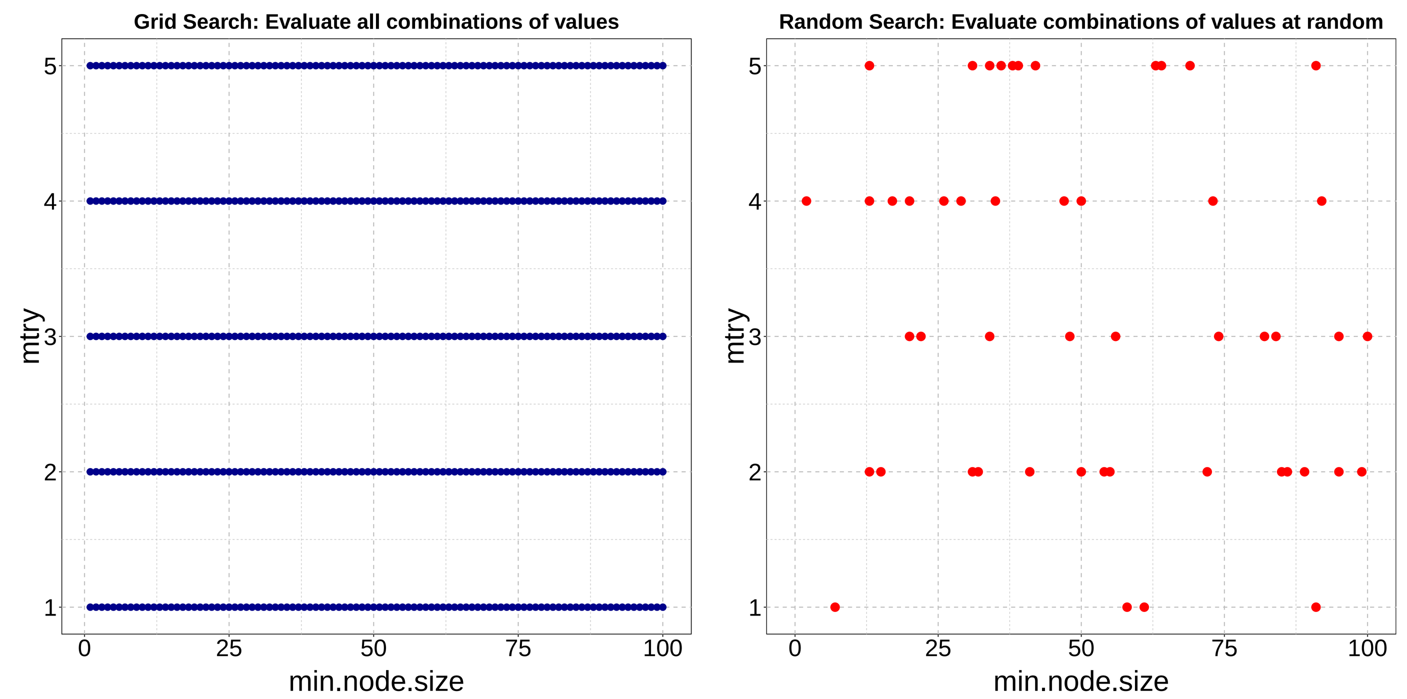
SMOTE is an over-sampling method in which the minority class is over-sampled by creating synthetic examples [2].




## HYPERPARAMETER OPTIMIZATION

**Question:** Can we improve the predictive performance of random forest by selecting good values for its hyperparameters [1]?

**Traditional methods** are grid and random searches, which may be ineffective or expensive for a large number of hyperparameters.

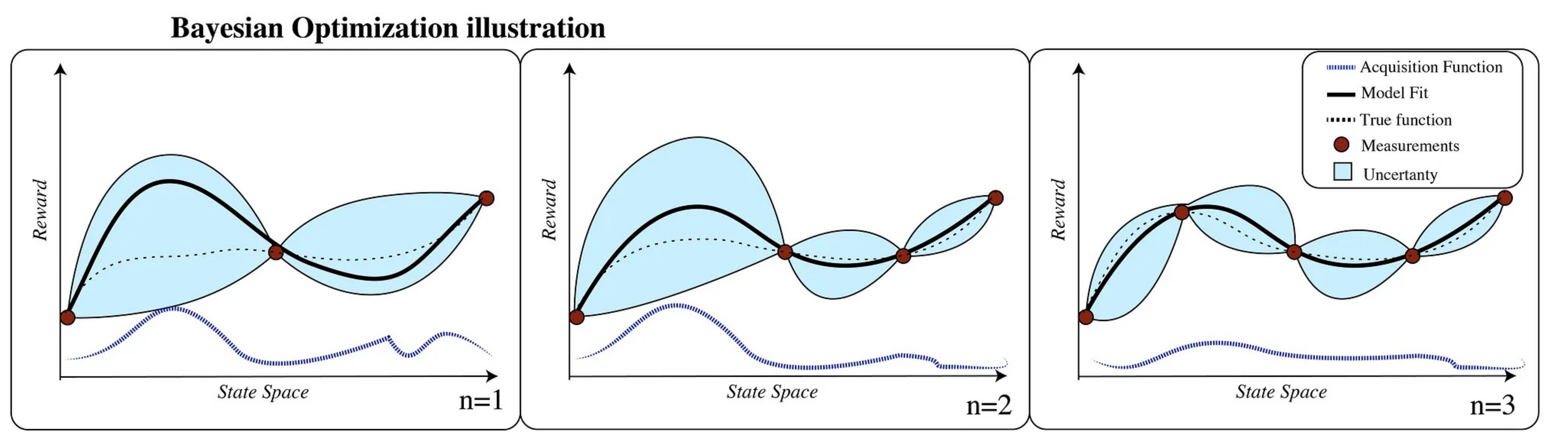



## BAYESIAN OPTIMIZATION

Bayesian Optimization is a global optimization method that is recently being used for hyperparameter optimization [1].

It has 4 steps:

1. Select an initial set of values using random sampling.
2. Fit a surrogate (*Gaussian process*) model that approximates the relation between the predictive performance and hyperparameters.
3. Maximize an acquisition function (*expected improvement*) to find the next combination of values to test.
4. Iterate between 2 and 3 until a maximum number of iterations.




Source: https://towardsdatascience.com/design-optimization-with-ax-in-python-957b1fec776f

## NUMERICAL COMPARISONS

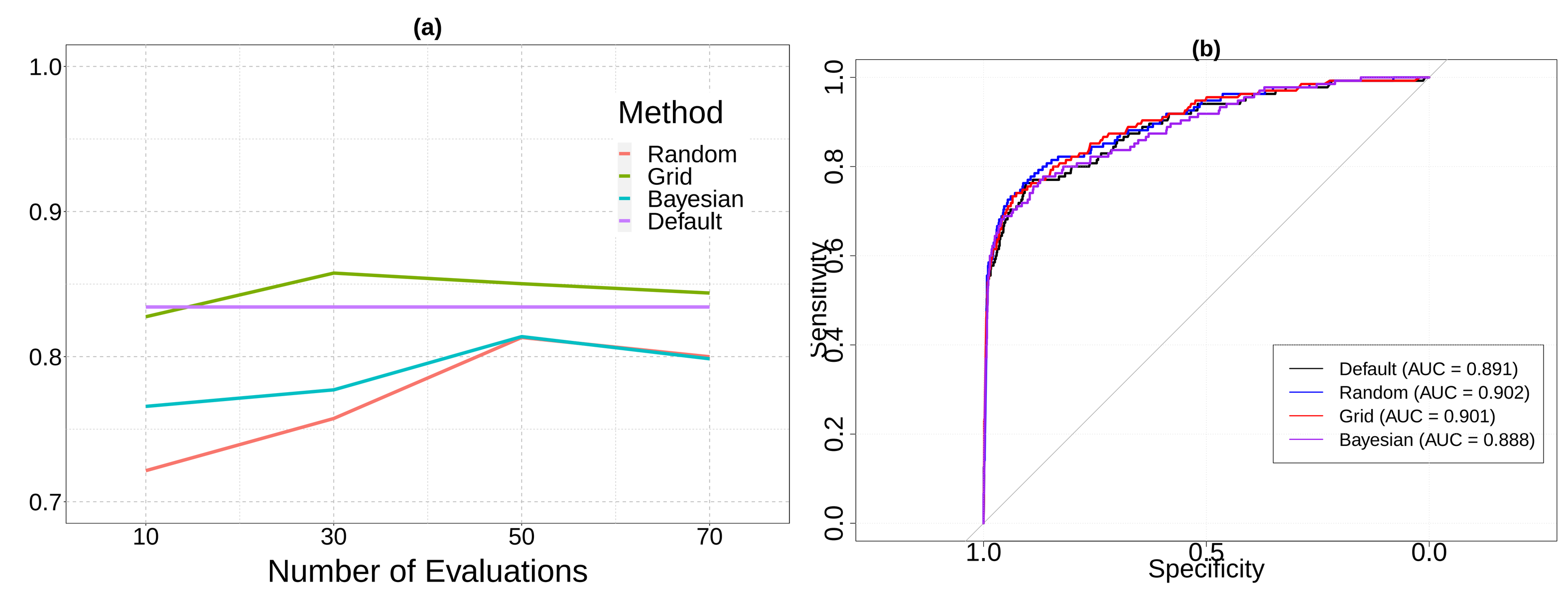
### Performance metrics

- Probability that a non-functioning pump is correctly identified among all non-functioning pumps (***sensitivity***).
- Probability that a functioning pump is correctly identified among all functioning pumps (***specificity***).

### Results

(a) Mean 5-fold cross-validation sensitivity estimate of 5 repetitions.

(b) Sensitivity and specificity obtained from test data with 1,968 functioning and 135 non-functioning pumps.




## DISCUSSION

- Hyperparameter optimization improves random forest.
- However, the improvement is marginal, suggesting that the algorithm is not tunable [5]. Moreover, standard Bayesian Optimization does not outperform the traditional methods.
- The performance of Bayesian Optimization may be improved by tuning its *meta*-hyperparameters, thus hindering the original problem.

**Acknowledgements**

The authors express their gratitude to the funding provided to support this study from National Science Foundation (NSF) EPSCoR RII Track-2 Program under the grant number OIA-2119691.

## References

[1] Bernd Bischl, Martin Binder, Michel Lang, Tobias Pielok, Jakob Richter, Stefan Coors, Janek Thomas, Theresa Ullmann, Marc Becker, Anne-Laure Boulesteix, Difan Deng, and Marius Lindauer. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Mining and Knowledge Discovery*, 13:e1484, 3 2023.

[2] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, jun 2002.

[3] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.

[4] Philipp Probst and Anne-Laure Boulesteix. To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research*, 18:1–18, 2018.

[5] Philipp Probst, Anne-Laure Boulesteix, and Bernd Bischl. Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, 20:1–32, 2019.

[6] Daniel L. Wilson, Jeremy R. Coyle, and Evan A. Thomas. Ensemble machine learning and forecasting can achieve 99% uptime for rural handpumps. *PLOS ONE*, 12(11):1–13, 11 2017.