Branching Out Carefully: A Cautious Take on Random Forests

David Feldman
Johns Hopkins University
dfeldm17@jhu.edu

Jarrett Huddleston
Johns Hopkins University
jhuddle6@jh.edu

Tyler Skow
Johns Hopkins University
tskow1@jhu.edu

Abstract

Methods that allow for principled cautious classification are essential for robust Artificial Intelligence (AI) systems that make consequential decisions. Our report provides a comprehensive survey of methods to implement cautiousness in Random Forest Classifiers. Our results demonstrate the ability to achieve accuracy surpassing baseline single-set performance while maintaining a modest abstention trade off. Furthermore, our experiments offer direction on which methods might be more suitable in various decision making contexts, including tasks that favor abstention over accuracy or vice versa. Code available at https://github.com/tskow99/cautious-random-forest.

1 Introduction

What if making a false positive classification is particularly costly for a given machine learning task? Perhaps the classifier's outputs influence medical diagnoses, drug prescriptions, or even parole decisions. In such instances, we may prefer a cautious classifier that takes into consideration the confidence of prediction given the severity of its consequences. Our project explores how cautiousness can be factored into Random Forest classifiers in a binary classification setting. Namely, we explore different algorithms for balancing confidence and abstention to yield cautious decision-making.

Before venturing further, we endeavor to provide a *precise* definitions of 'cautiousness', 'uncertainty', 'cautious classifier' and 'imprecise classification' to avoid ambiguity in our methods. Unsurprisingly, there are a variety of distinct definitions for these terms across the literature related to machine learning classification (Lakshminarayanan et al. (2017), Smith et al. (2018)). We differ discussion on the merits of various permutations of these terms to other forums, and adopt the following for brevity:

- **Cautiousness:** The ability to refrain from making a decision given a certain threshold or in the absence of key information.
- **Uncertainty:** The extent to which our models decisions exist close to a decision boundary, or, the data available for training makes classifying such a point ambiguous. We use several mathematical definitions to measure uncertainty in our methods, including:
  - **Class Probability Greater Than a Threshold:**

    $$P(y = c \mid x) > \tau$$

    Where:
    - $P(y = c \mid x)$ is the probability of class c given input x.
    - τ is the predefined threshold.
  - **p-Value for Non-conformity Score Greater Than Threshold α:**

    $$p > \alpha$$

    Where:

Preprint. Under review.

    - p is the p-value corresponding to the non-conformity score.
    - α is the predefined significance threshold.
- **Cautious Classifier:** A type of classifier that includes mechanisms to abstain from predictions when confidence is low.
- **Imprecise Classification:** A scenario where a classifier outputs a prediction that spans class boundaries.

As a further point of clarification, because we are operating in a binary setting, we make no distinction between an 'abstention' from making a decision and generating an imprecise classification where an interval of classes is returned. In a multi-class setting, when an imprecise classification can yield distinct intervals, such a distinction makes sense, but not in our setting.

Our work is the first of its kind to provide a comprehensive survey of the performance of various cautious and imprecise Random Forest (RF) classifiers across multiple binary-labeled datasets. While prior research, such as Zhang et al. (2023), developed methods for cautious classification in RFs by focusing on tree weighting and voting schemes, our study builds upon their contributions by broadening the scope of cautious implementations evaluated. Specifically, we use similar datasets and evaluation metrics and include their methods as part of our benchmarked algorithms. However, we extend this line of research by introducing additional approaches, such as conformal prediction and RFs with fuzzing. Our results demonstrate the ability to achieve accuracy surpassing baseline single-set performance while maintaining modest abstention trade-offs. These results may be promising for practitioners who assess the determinism-accuracy trade off our methods offer may be worthwhile for their specific classification task.

2 Methods

In this project, we compared the performance of six implementations of random forest:

- **Baseline Random Forest Model**
  - Our baseline random forest was built with Sci-kit Learn Pedregosa et al. (2011). In short, a basic random forest classifier works by drawing bootstrap samples from our original data observations, learning a decision tree (after picking a subset of the features of the data) for each bootstrap sample, and aggregating the results at the end via majority vote for classification. We also used grid search cross-validation to select optimal hyper parameters and predicting the test error with 10-fold cross-validation.
- **A naïve cautious classifier**
  - Our naïve cautious classifier, which was based on our baseline random forest model, removed predictions where confidence (measured as the probability for the favored class) was below a user-defined threshold. This model also included hyper parameter tuning via grid search cross-validation in a similar fashion to the vanilla random forest. This classifier was inspired by the work from Ferri and Hernandez-Orallo (2004)
- **A naïve cautious classifier with class specific thresholds**
  - As an extension to the above, rather than exclusively considering a global threshold, we fit class-specific thresholds. The motivation for class-specific thresholds is to account for classes with distinct probability distributions, e.g. as a results of class imbalance in a dataset Ferri and Hernandez-Orallo (2004). Fitting a single threshold to two unequally represented classes may not yield good results. To allow for these thresholds to be truly distinct, class label selection is done by

    $$\text{pred} = \text{argmax} \frac{p_i}{t_i}$$

    where p_i is the probability of class i and t_i is the threshold for class i. We fit our class specific thresholds with a simple grid search using u65 score for evaluation. This method is described in Ferri and Hernandez-Orallo (2004).
- **Cautious Weighted Random Forest**

  - As detailed in Zhang et al. (2023), the cautious weighted random forest first fits a plain random forest model, calculates interval-valued probability estimates, and uses these estimates to calculate measures of belief and plausibility, which quantify whether an instance belongs to class 1 (calculation of belief and plausibility also requires calculating optimal weights for each tree in the random forest, which are found via a convex optimization solver). If an observation has belief of at least 0.5, the cautious random forest predicts a 1, if the plausibility is below 0.5, it predicts a 0, and if neither occurs, it labels the point undetermined.
  - The pseudocode for the weighted cautious random forest is included in the appendix A. Similarly to the models discussed above, we utilized grid search cross-validation to select the optimal hyperparameter values.

- **Conformal Prediction**
  - This classifier generates prediction sets with a predefined coverage guarantee and labels observations as undetermined when these confidence intervals overlap. This classifier also supports nonconformity scores for robust decision making. Our classifier built upon the work from Dreyfus-Schmidt (2024)

- **Fuzzy Random Forest**
  - A fuzzy random forest utilizes the ability of a fuzzy decision tree to handle uncertainty in data and the efficacy of bagging to create a more accurate classifier where the data and labels may not be clearly delineated.
  - Fuzzy decision trees are a popular technique for evaluating data where certainty may not be present, and with it it we hope to draw a contrast between a classifier specializing in uncertainty and the cautious weighted random forest, which uses uncertainty to approach decisions cautiously. We build our fuzzy random forest on the work from Liu et al. (2022)

2.1 Evaluation Metrics

Standard evaluation metrics for classification, e.g accuracy, precision, recall, AUC and F1 are insufficient for our experiments. We require metrics that appropriately reward/penalize based on both a model's accuracy and abstention rate. For example, a model that is only confident enough to make predictions for 1% of the test dataset may not be useful, even if it achieves 100% accuracy. To illustrate the abstention vs. accuracy trade off, we collect single-set accuracy, determinacy, abstention rate and u65 score for all classifiers. Derivation of these metrics were adapted from Zhang et al. (2023). The u65 score balances reward for abstention and prediction by rewarding 0.65 when the classifier does not make a prediction. Detailed definitions of each of these metrics are included below:

Single-Set Accuracy

$$\text{Single-Set Accuracy} = \frac{\sum_{i \in D}(y_i = \hat{y}_i)}{\text{Single-Set Length}}$$

- The accuracy of the set of points the classifier made a decision on.

Determinacy

$$\text{Determinacy} = \frac{\text{Single-Set Length}}{\text{Total Number of Instances}} \times 100$$

- Determinacy is the proportion of observations where the model provides a determinate prediction
- Single-Set Length: Number of determinate instances.

Abstention

$$\text{Abstention} = 100 - \text{Determinacy}$$

- Proportion of the test set on which the classifier did not make a decision.

U65 Score

$$\text{U65 Score} = 65 + \frac{(\text{Single-Set Accuracy} - 65) \cdot \text{Determinacy}}{100}$$

- A discounted accuracy measure that accounts for both accuracy and determinacy.

For our baseline random forest model, our main evaluation metric was the accuracy score. Future avenues of research might explore additional metrics for comparing cautious classifiers to non-cautious classifiers.

3 Data Processing and Cleaning

For this project we ran our random forest models on 4 different datasets. We used the COMPAS Recidivism dataset (COMPAS data), which was the basis for a biased algorithm that resulted in unjust incarceration for reasons such as race and aims to predict two-year recidivism(Julia Angwin and ProPublica); the German Credit dataset(Hofmann (1994)), which aims to predict whether a person will be a good or bad creditor; and two medical datasets concerning heart disease(Janosi and Detrano (1989)) and breast cancer(Wolberg and Street (1993)), both of which aim to predict whether or not a patient has the respective disease. Although these data sets are from different domains and have different underlying properties, they are similar in that cautious prediction could potentially be beneficial over binary classification, as incorrect classification of the target variables for each of these datasets could result in unfair imprisonment, loss of money, or death, and thus caution could be advantageous. As a result, we utilized our various random forest classifiers on these datasets in order to compare how our methods perform on different datasets.

Each dataset we included in our analysis required its own pre-processing, as each dataset had different structures and required different amounts of work prior to feeding it into our models.

For the German Credit Data, pre-processing included one-hot encoding categorical features and dropping the original categorical features. This dataset did not have any missing data, unlike some of the other datasets we worked with, as discussed below.

Since the heart disease dataset included many instances of missing data, we built a missing data model for each variable that had missing elements (7 variables in total). That is, we fit 7 different regression models (predicting whether a patient has heart disease) that included all variables that were fully realized and one of the missing data indicators, and analyzed their output. If the coefficient for the missing data indicator was significant (on a 5% level), this meant that the missingness status of the variable has a significant effect on the value of our target variable. As a result, our prediction for a patient having heart disease would depend on whether or not this variable is missing, so imputation would be inappropriate. After running these 7 regressions, we found that 3 of these variables had significant coefficients (their coefficients had p-values below 5%) and therefore imputation and other techniques to fill in missing data would not be appropriate for these variables. For the remaining 5 variables that were not significant, we utilized Sci-kit Learn's iterative imputer (which does imputation in an iterative round-robin fashion) to fill in the missing values. Because the goal of our project was not to deliver the most accurate predictions for each dataset, but instead to compare measures of cautiousness across the different datasets (and because of the short time frame to complete our project), we decided to drop these significant variables from our data. This dataset also required one-hot encoding for categorical variables.

The COMPAS dataset had only one column with missing values, and after performing a missing data analysis similar to the one done on the heart disease data, we found that this variable's coefficient was significant, and for the same reasons mentioned above we dropped it from the dataset. We also had to drop other columns from this dataset that were not relevant for prediction (such as name and date of birth).

Since the breast cancer dataset didn't have any missing values, no additional data processing was required outside of separating the target variable from the predictors.

Lastly, each dataset was split into training, testing, and calibration datasets (for our conformal prediction), with testing size equal to one third of the data size, and the calibration dataset being one tenth the size of the remaining two thirds of data (or one fifteenth the size of the overall data).

4 Experiments

4.1 Implementation Details

All of the code for our experiments was written and tested in python 3.9. Each classifier was implemented as follows:

- **Baseline Random Forest**: we used the random forest implementation provided by the python scikit-learn package.
- **Naive cautious classifier**: our implementation is derived from the approach outlined in Ferri and Hernandez-Orallo (2004).
- **Naive cautious classifier with class specific thresholds**: our implementation is derived from the approach outlined in Ferri and Hernandez-Orallo (2004).
- **Conformal Prediction**: the implementation is based off of the code provided in Dreyfus-Schmidt (2024)
- **Cautious Weighted Random Forest**: We used the implementation provided in Zhang et al. (2023).
- **Fuzzy Random Forest**: Our implementation was built off of the codebase in Liu et al. (2022) and adapted to fit within our pipeline.

To evaluate each of our datasets, we followed the approach outlined in §3.

4.2 Results

Table 1: Baseline Random Forest Results

Dataset	Accuracy (%)
German Credit	74.00
Heart Disease	86.18
Breast Cancer	95.74
COMPAS	66.39

Our results are summarized in the tables 1 and 2. Table 1 displays our baseline random forest results, and we see that it performs best on the breast cancer data and performs worst on the COMPAS data. We provide these results in an effort to contextualize the diversity in data. On one end of the spectrum we have a challenging dataset to classify in COMPAS, achieving slightly better then random, and on the other end, Breast Cancer is classified with nearly perfect accuracy. For a more detailed review on how the baseline classifiers performed on each dataset from a confidence perspective, see B.

Tables 2 displays our different cautious classifiers' performance on our four datasets. In particular, we see that of our cautious classifiers, the conformal classifier has the lowest determinacy of our predictors but delivers the highest single set accuracies for each dataset except the Breast Cancer dataset.

Additionally, the weighted cautious random forest has the highest determinacy of the cautious classifiers for each dataset, yet has the lowest single set accuracy for each dataset aside from (again) the breast cancer dataset, above the fuzzy random forest. This is a clear example of the tradeoffs our classifiers face in terms of balancing high predictive accuracy for a small set of observations and delivering determined predictions. 5 further illustrates this point. In the case of this dataset, the trade off is elbow shaped, suggesting sharp gains by abstaining from the least determinant test observations. This chart was created by changing our confidence threshold from 0.5-0.99. Additional figures showing the abstention-accuracy trade off for each dataset are available in the appendix: C.

5 Discussion

We can observe from Table 2 that the cautious weighted random forest performs as well or better than the other random forest classifiers we examined with respect to the U65 score. In three of the

Table 2: Evaluation Results for RandomForest, ConformalPredictor, CWRF, NaiveCautiousClassifier, NaiveCautiousClassifier with class specific thresholds (CSTCautiousClassifier) and FuzzyRandomForest on four datasets: German Credit, Heart Disease, Breast Cancer and COMPAS. Metrics defined in 2.1

Classifier	Dataset	U65 Score	Single Set Accuracy	Determinacy	Precise Accuracy	Abstention
RandomForest	german_credit_data	74.0	74.0	100.0	74.0	0.0
ConformalPredictor	german_credit_data	74.52	87.95	41.5	74.0	58.0
CWRF	german_credit_data	76.0	76.0	100.0	78.0	0.0
NaiveCautiousClassifier	german_credit_data	77.25	87.27	55.0	74.0	45.0
CSTCautiousClassifier	german_credit_data	75.48	86.6	48.5	74.0	51.5
FuzzyRandomForest	german_credit_data	70.0	70.0	100.0	70.0	0.0
RandomForest	heart_disease_data	86.18	86.18	100.0	86.18	0.0
ConformalPredictor	heart_disease_data	86.96	97.56	67.43	86.18	32.57
CWRF	heart_disease_data	87.5	87.5	100.0	87.15	0.0
NaiveCautiousClassifier	heart_disease_data	88.88	95.76	77.63	86.18	22.37
CSTCautiousClassifier	heart_disease_data	86.18	86.18	100	86.18	0.0
FuzzyRandomForest	heart_disease_data	69.08	69.08	100.0	69.08	0.0
RandomForest	breast_cancer_data	95.74	95.74	100.0	95.74	0.0
ConformalPredictor	breast_cancer_data	75.18	92.75	36.7	95.74	63.3
CWRF	breast_cancer_data	96.81	96.81	100.0	96.81	0.0
NaiveCautiousClassifier	breast_cancer_data	96.31	99.42	90.96	95.74	9.04
CSTCautiousClassifier	breast_cancer_data	95.74	95.74	100.0	95.74	0.0
FuzzyRandomForest	breast_cancer_data	92.02	92.02	100.0	92.02	0.0
RandomForest	compas_data	66.6	66.6	100.0	66.6	0.0
ConformalPredictor	compas_data	67.85	77.92	22.06	66.6	77.94
CWRF	compas_data	64.5	64.37	78.91	63.6	21.09
NaiveCautiousClassifier	compas_data	69.8	77.84	37.37	66.6	62.63
CSTCautiousClassifier	compas_data	70.25	75.71	48.99	66.6	51.01
FuzzyRandomForest	compas_data	59.01	59.01	100.0	59.01	0.0

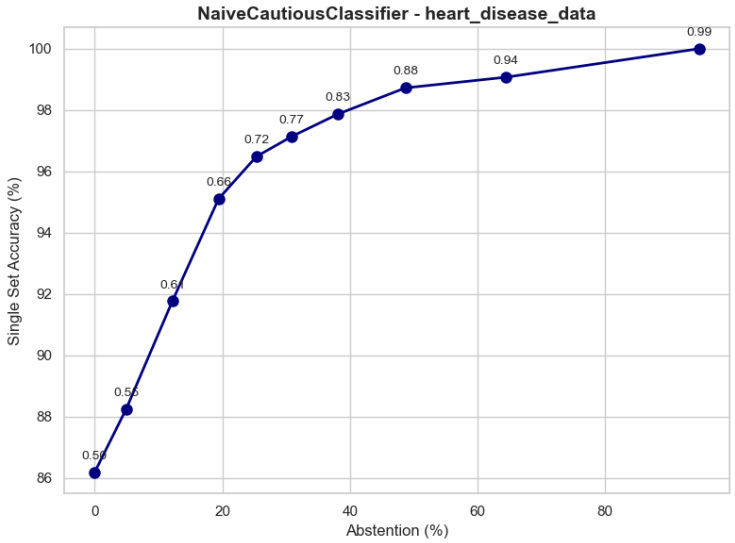


Figure 1: Accuracy vs. Abstention Tradeoff for Breast Cancer Dataset

four datasets, the CWRF is as good or better than the standard random forest implementation, and in the one case where it underperforms it is only by 2%. Interestingly, the CWRF only performs worse

than the standard random forest in the case where it is unable to provide a decision for 20.39% of samples in the test dataset.

The CWRF also consistently outperforms the fuzzy random forest, another model designed to handle uncertainty in the data. The key distinction between the two is that CWRF considers the margin by which a decision is made, and the fuzzy random forest instead considers uncertainty in data in the component trees. Like the standard random forest, the only case in which the fuzzy random forest yields a better accuracy than the CWRF is with the COMPAS dataset, where it is able to make a prediction for every sample.

The CWRF is also able to provide decisions more often than the other three cautious classifiers. As noted in §4.2, the conformal classifier has the lowest determinacy of any of the evaluated predictors, 41.65% on average across the four datasets, but in $\frac{3}{4}$ of the cases it provides the highest single set accuracy. The naive cautious classifier, while not as accurate as the conformal predictor, is still able to consistently achieve a higher accuracy score than the CWRF, however with an average determinacy of 65.21%. The CWRF has an average determinacy of 94.99%. These discrepencies highlight the significance of the U65 score, as the conformal predictor and naive cautious classifier seem to achieve their relatively high accuracies by considering fewer samples: the only ones they can confidently classify. With the U65 score we can evaluate a broader "usefulness" of each predictor, as it captures how much of the dataset is able to be classified as well as the accuracy. After all, a predictor is much less useful if it can only make predictions on less than half of the samples. Considering the U65 score, the usefulness of the CWRF becomes more apparent. As evidenced in Table 2, the CWRF's U65 score generally matches or beats that of the other classifiers. Therefore, it can be considered similarly useful because of its ability to make cautious predictions for a higher proportion of the dataset, which is significant in cases where you may not be able to use a tool that classifies only a small portion of the data. Also of note is that using class-specific thresholds does not appear to improve performance across any of the datasets. In the heart disease and breast cancer datasets, the class-specific thresholded classifiers perform the same as the baseline model, while in the COMPAS and german credit datasets there is little distinction from the naïve cautious classifier. It could be that in a binary setting, generating class-specific thresholds has little utility.

5.1 Comparison to Prior Work

We are able to match the performance of the CWRF with the original paper, Zhang et al. (2023). Two of our datasets were used in the original CWRF paper (the breast cancer and heart disease datasets), and in both we achieve similar scores. Comparing our observed U65 scores for the WCRF with those of the original paper for the two datasets in common, we find that the largest difference is 1.12 (84.87 in our experiment vs 83.75 in the original paper) for the heart disease dataset. This difference is small enough that it can be attributed, at least in part, to the difference in tuning of the hyperparameters.

5.2 Limitations and Future Work

One significant limitation we faced in our work was the time constraints for the project. It is computationally challenging to find the optimal set of hyperparameters for each of our classifiers. Given more time to tune the hyperparameters, we could be more certain in the performance of each of these models.

In addition, we would also try to incorporate the variables dropped due to significant coefficients from our missing data model into our analysis in the future, which would potentially involve making more assumptions on our data, enlarging the parameter space, or changing our parameters to something that is easier to estimate.

Lastly, we would also like to explore altering our classifiers to handle multi-class problems instead of just binary classification. This would allow our models to handle a much wider array of problems and make our methods as general as possible.

6 Conclusion

Our work is a first of its kind survey of the methods used to consider uncertainty and caution in random forests. We examined six random forest implementations: a standard/baseline random

forest, a conformal predictor, a naive cautious classifier, a naive cautious classifier with class-specific thresholds, a fuzzy random forest, and a cautious weighted random forest, across four different datasets. In our analysis we were able to identify trends in the five models, such as the conformal predictor and the naive cautious classifier achieving higher single-set accuracy while having significantly lower determinacies. We also identified the usefulness of the cautious weighted random forest: its ability to make cautious classifications while being able to maintain a notably higher determinacy. Ultimately, our results demonstrate the ability to achieve accuracy that surpasses baseline single-set performance while maintaining modest abstention trade-offs, which we hope to be promising for practitioners addressing classification problems for which caution with low abstention is a requirement.

A Cautious Weighted Random Forest pseudo-code

Input: random forest RF, tree weights w_t, IDM parameter s, set of test instances X
Output: predictions $\hat{Y}$ for test instances

$\hat{Y} \leftarrow \{\}$
for $x_i \in X$ **do**
  for $f_t \in RF$ **do**
    Compute $I_t(x_i)$ via Eq. (11)
  Calculate bel_i^1 via Eq. (12)
  Calculate pl_i^1 via Eq. (13)
  if $bel_i^1 \geq 0.5$ **then**
    $\hat{y}_i \leftarrow 1$
  else if $pl_i^1 < 0.5$ **then**
    $\hat{y}_i \leftarrow 0$
  else
    $\hat{y}_i \leftarrow \{0, 1\}$
  $\hat{Y} \leftarrow \hat{Y} \cup \hat{y}_i$

Figure 2: Cautious Weighted Random Forest pseudo-code fron Zhang et al. (2023)

B Baseline Confidence Across Datasets

Figure 3a, Figure 3b, Figure 3c, and Figure 3d demonstrate the baseline confidence across our four datasets.

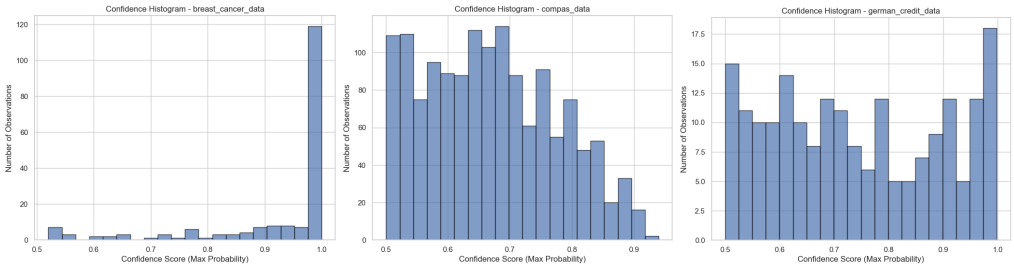
C Accuracy-Abstention Trade off by dataset

Figure 4, Figure 5, and Figure 6 demonstrate the accuracy-abstention trade off by dataset.

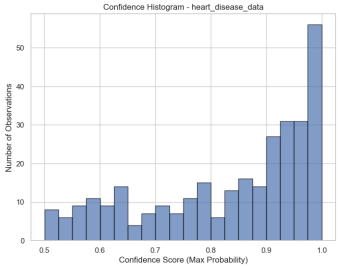
References

COMPAS data. [link].

Leo Dreyfus-Schmidt. 2024. Measuring model's uncertainty with conformal prediction.



(a) Breast Cancer Dataset (b) COMPAS Dataset (c) German Credit Dataset



(d) Heart Attack Dataset

Figure 3: Classifier confidence on test data for each dataset

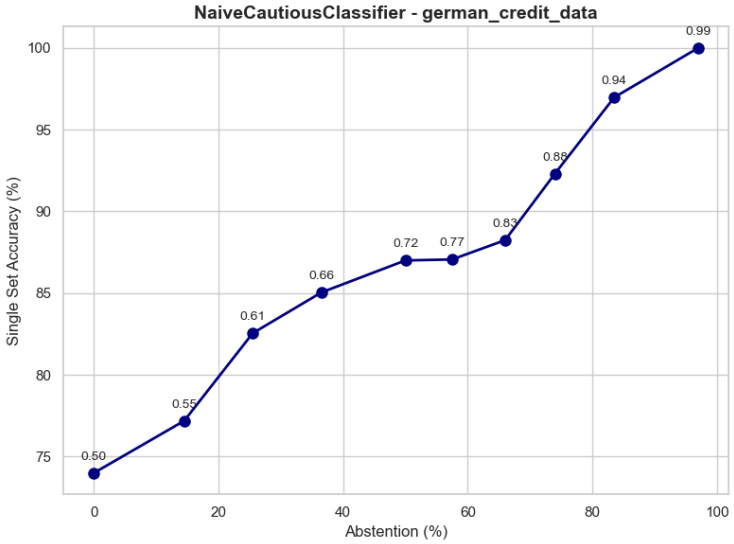


Figure 4: Accuracy vs. Abstention Trade off for German Credit Data

Cèsar Ferri and Jose Hernandez-Orallo. 2004. Cautious classifiers. pages 27–36.

Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.

Steinbrunn William Pfisterer Matthias Janosi, Andras and Robert Detrano. 1989. Heart Disease. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C52P4X.

Surya Mattu Lauren Kirchner Julia Angwin, Jeff Larson and ProPublica. [link].

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6402–6413.

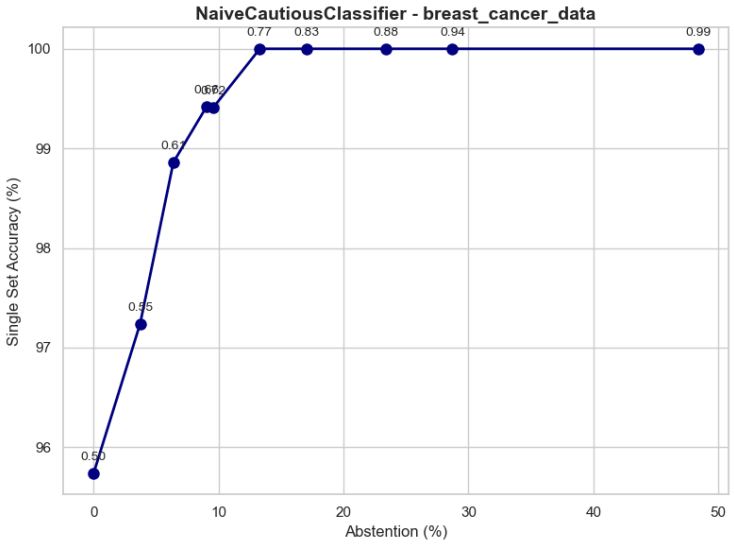


Figure 5: Accuracy vs. Abstention Trade off for Breast Cancer Data

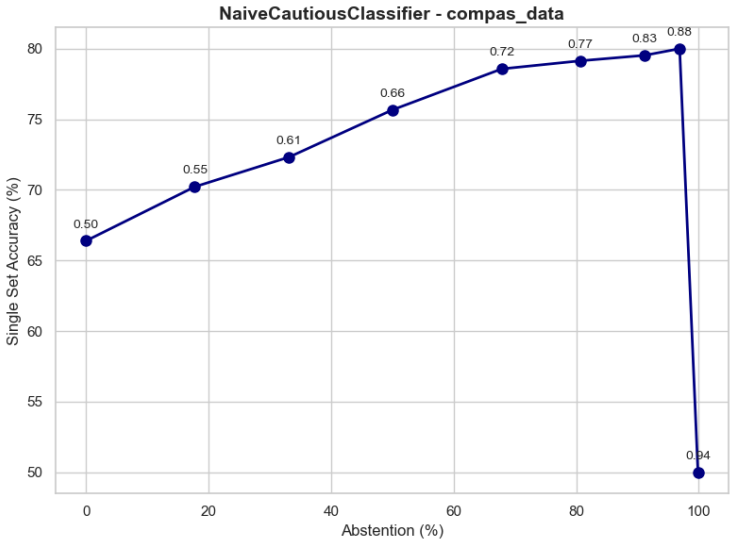


Figure 6: Accuracy vs. Abstention Trade off for COMPAS Dataset

Zhaoqing Liu, Anjin Liu, Guangquan Zhang, and Jie Lu. 2022. An empirical study of fuzzy decision tree for gradient boosting ensemble. In *AI 2021: Advances in Artificial Intelligence*, pages 716–727, Cham. Springer International Publishing.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Lewis Smith, Yarin Gal, and Zoubin Gharamani. 2018. Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*.

Mangasarian Olvi Street Nick Wolberg, William and W. Street. 1993. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5DW2B.

Haifei Zhang, Benjamin Quost, and Marie-Hélène Masson. 2023. Cautious weighted random forests. *Expert Systems with Applications*, 213:118883.