Towards Understanding Fairness and its Composition in Ensemble Machine Learning

Usman Gohar  
Dept. of Computer Science  
Iowa State University  
Ames, IA, USA  
ugohar@iastate.edu

Sumon Biswas  
School of Computer Science  
Carnegie Mellon University  
Pittsburgh, PA, USA  
sumonb@cs.cmu.edu

Hridesh Rajan  
Dept. of Computer Science  
Iowa State University  
Ames, IA, USA  
hridesh@iastate.edu

Abstract—Machine Learning (ML) software has been widely adopted in modern society, with reported fairness implications for minority groups based on race, sex, age, etc. Many recent works have proposed methods to measure and mitigate algorithmic bias in ML models. The existing approaches focus on single classifier-based ML models. However, real-world ML models are often composed of multiple independent or dependent learners in an ensemble (e.g., Random Forest), where the fairness composes in a non-trivial way. How does fairness compose in ensembles? What are the fairness impacts of the learners on the ultimate fairness of the ensemble? Can fair learners result in an unfair ensemble? Furthermore, studies have shown that hyperparameters influence the fairness of ML models. Ensemble hyperparameters are more complex since they affect how learners are combined in different categories of ensembles. Understanding the impact of ensemble hyperparameters on fairness will help programmers design fair ensembles. Today, we do not understand these fully for different ensemble algorithms. In this paper, we comprehensively study popular real-world ensembles: bagging, boosting, stacking and voting. We have developed a benchmark of 168 ensemble models collected from Kaggle on four popular fairness datasets. We use existing fairness metrics to understand the composition of fairness. Our results show that ensembles can be designed to be fairer without using mitigation techniques. We also identify the interplay between fairness composition and data characteristics to guide fair ensemble design. Finally, our benchmark can be leveraged for further research on fair ensembles. To the best of our knowledge, this is one of the first and largest studies on fairness composition in ensembles yet presented in the literature.

Index Terms—fairness, ensemble, machine learning, models

I. INTRODUCTION

Machine learning (ML) is ubiquitous in modern software today. Due to the black-box nature of ML algorithms and its applications in critical decision-making, fairness in ML software has become a huge concern. Measuring ML fairness and mitigating the discrimination has been studied extensively. Recent work in software engineering has shown the need to produce fair software and detect bias in complex ML software environments. Prior research has mostly focused on fairness in standalone classifiers (e.g., Logistic Regression, SVM) and better generalization on unseen data, especially in smaller datasets. Ensemble models combine the predictions of multiple base learners to make the final prediction, e.g., Random Forest uses a large number of decision trees, with the majority class the final output. Ensemble models are the most mentioned ML algorithms on Kaggle and in previous SE works on fairness, ensemble models comprise more than 80% of the total models. Like traditional ML models, ensemble models can also suffer from unfairness problem that discriminates against population subgroups based on race, gender, etc. Although many fairness mitigation techniques exist, they do not always generalize well. Therefore, if we better understand the fairness composition in ensembles, we can design fair ensemble models without applying mitigation techniques. In this paper, we have conducted an empirical study to understand the composition of fairness in ensembles and the interplay of their properties with fairness.

Recently, multiple works have shown that ensembles can be leveraged to enhance fairness and mitigate bias in ML models. Grgic-Hlaca et al. first explored fairness properties of random selection ensemble, only theoretically. Bower et al. explored how fairness propagates through a multi-stage decision process like hiring. Similarly, Dwork et al. introduced a framework to understand the composition of fairness in ensembles that only utilize AND, OR operators to make a decision, e.g., two credit bureaus’ (AND) report a score to determine loan eligibility. Feffer et al. studied how ensembles and bias mitigators can be combined using modularity to improve stability in bias mitigation. Therefore, it is evident that fairness in ensembles and their composition is non-trivial. Moreover, prior works in SE have shown the impact of training processes such as hyperparameter optimization, data transformation, etc., on the fairness of ML software. We postulate that ensemble hyperparameters also impact unfairness in ensembles, and failure to study them can amplify bias. However, ensemble hyperparameters are different than typical ML model hyperparameters as they dictate the design of the ensemble, e.g., number of learners, learning method, etc. However, no empirical study has been conducted to understand fairness composition in ensembles and the effect of their hyperparameter space on fairness. To this end, we have created a benchmark of 168 real-world ensemble models from...
Kaggle and designed experiments to measure their fairness. We analyze fairness composition in ensemble criteria such as parallel and sequential ensembles, homogeneity of models, and different ensemble methods such as bagging, boosting, voting, stacking, etc., and all the ensemble classifiers available in the popular Scikit-learn [33]. Specifically, we answer the following overarching research questions:

**RQ1:** What are the fairness measures of various ensemble techniques? **RQ2:** How does fairness compose in ensemble models? **RQ3:** Can ensemble-related hyperparameters be chosen to design fair ensemble models?

To the best of our knowledge, this is the first work to experimentally evaluate the fairness composition in popular ensembles and elicit fair ensemble design considerations. Our results show that fairness in ensembles composes in the base learners and fair ensemble models can be built by carefully considering the composition. The analyses show which learners cause fairness problems which software developers can leverage to develop frameworks to measure fairness in base learners to encourage transparency. We also identify and explore ensemble-related hyperparameters to design fair ML models for each ensemble type. Lastly, we provide a comprehensive review of fairness composition in ensembles that will help direct future research in the area. Overall, the following are the key contributions of this paper:

- Explored fairness composition and its interplay with data characteristics and individual learners to mitigate bias.
- Empirically evaluates fairness patterns of popular ML ensemble models.
- We identified ensemble design considerations and hyperparameters that would guide developers in fair ensemble design and mitigate inherent unfairness effectively.
- A comprehensive fairness benchmark of popular ensembles that leveraged for further research on building fairness-aware ensembles. The benchmark, code and experimental results are available: https://github.com/UsmanGohar/FairEnsemble

The rest of the paper is organized as follows: §II describes the motivation and background on ensembles. In §III we describe the methodology for our study, including benchmark collection, datasets and fairness and accuracy measures used, and finally, the experiment setup & design. In §IV we discuss the state of fairness in ensembles (RQ1) and how it composes (RQ2). §V discusses the design criteria to improve fairness in ensembles (RQ3). Finally, we discuss the implications of our work in §VI, threats to validity in §VII related works in §VIII and then present the conclusion in §IX.

### II. Motivation and Background

In this section, we use a motivating example to illustrate the complexity of fairness composition in ensembles and the need to study bias induced by certain ensemble parameters.

#### A. Motivating example

Ensemble models are widely deployed to win competitions in online communities like Kaggle due to their superior performances [16]–[20], [23]. In prior SE works [12], [13] on fairness, more than 80% of the models were ensemble based. However, those works did not consider fairness composition of individual learners, its effect on fairness of ensembles and the inherent bias in ensemble methods, which is non-trivial. Hence, not studying fairness composition in ensembles fails to capture the complete fairness of an ML pipeline. Consider the code snippet below of a top-performing model (Voting ensemble) from Kaggle, which is used to predict the income of an individual (German Credit dataset).

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.ensemble import VotingClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import KNeighborsClassifier
from sklearn.ensemble import AdaBoostClassifier
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.naive_bayes import GaussianNB

models = []
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('RF', RandomForestClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC(gamma='auto')))#
models.append(('XGBM', XGBClassifier()))
models.append(('LGBM', LGBMClassifier()))
model = VotingClassifier(estimators=models, voting='soft')
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
```

Voting ensemble is a type of heterogeneous ensemble that combines the predictions of dissimilar learners. This ensemble comprises multiple learners (lines 2-10) and uses a voting mechanism (line 11) to make the prediction. In soft voting, class label (1 or 0) with the higher average probability from the learners is chosen as the final prediction. We found that this ensemble is biased towards female applicants (Protected attribute: Sex) in terms of statistical parity difference (SPD:-0.203). In this example, before training the ensemble, a developer must decide the number of learners, select which learners to use, and the voting type (soft/hard). However, we found that ML libraries do not provide any fairness recommendations to build ensembles. Do these learners introduce unfairness in the predictions? How does the number of learners impact the fairness of the ensemble? More importantly, we observed that individual learners have their own fairness measures when analyzed in isolation but might result in an unfair model when used in an ensemble. For instance, our analysis shows that dropping XGBClassifier and LGBMClassifier (lines 9-10) can improve fairness by 27% (SPD:-0.148). Interestingly, we discuss later how these two learners are inherently fair themselves and not responsible for the unfairness.

Furthermore, prior research has shown the impact of hyperparameters on fairness [12], [25], [32]. Ensemble hyperparameters dictate how ensembles combine learners for final prediction. In this example, if developers used "hard" voting (line 11), the fairness of the ensemble would improve (SPD: -0.195). Similarly, some of these hyperparameters also affect the design properties of the learners which impacts fairness. XGBoost (line 9) is another example of ensemble (boosting). Boosting builds an ensemble of trees (learners) using various methods. What properties of these trees (e.g. tree depth, number of features etc.) and the learning method impact the overall fairness of a boosting ensemble? Exploring these parameters will help developers understand how to design fair ensembles.
Therefore, in addition to understanding fairness composition in the learners, it is equally important to understand how the design of ensembles using these parameters impacts fairness.

B. Ensemble learning in ML software

**Ensemble models** are a class of ML classifiers where the predictions from different learners (models) are pooled using a combination method (voting, average, random, etc.) to make the final predictions. In the motivational example above, we only discussed a single type of ensemble. Categories of ensembles are based on homogeneity, learning technique, and ensemble types. All the ensemble types covered in our study and the corresponding classifiers are given in Table I. There are mainly two categories of ensembles: Sequential and Parallel.

**Sequential Ensembles.** These ensembles sequentially generate base learners. Each learner in the ensemble depends on the previous learners in the sequence because the next learner attempts to correct the wrong predictions from the previous learner and so on [34]. AdaBoost is an example of a sequential model where it reweights (higher) misclassified examples.

**Parallel Ensembles.** Parallel ensembles train individual base learners in parallel and independent of each other. These learner are combined using techniques such as bagging (random sample of data with replacement) or voting, which encourages improved variance [34] e.g., Random Forest.

**Homogeneity of ensembles.** Ensemble methods that use single-type base learners are called homogeneous models [35]. These individual learners are combined to generate the final result, e.g., XGBoost and AdaBoost use decision trees. By contrast, heterogeneous ensembles combine the predictions of dissimilar individual learners [35]. A popular heterogeneous ensemble method is Voting. Finally, ensemble method types are divided into Boosting, Bagging, Voting, & Stacking [33].

### III. Methodology

In this section, we discuss the benchmark collection process, the datasets, and fairness and accuracy measures. Finally, we describe the experimental design and setup.

A. Benchmark Collection

For our experiments, we collected ensemble models from Kaggle [36] for datasets that have been used in prior fairness literature [12], [28], [37]. Unlike these works, we only collect ensemble-based models for evaluation. Specifically, we collect all ensemble classifiers available via the popular sci-kit learn library [33]. We follow a similar benchmark collection process as in [13]. Table I summarizes the datasets and the classifiers in the benchmark.

For each dataset, we collected Kaggle kernels for each ensemble type in Table I and classifiers available in scikit-learn. We filter these kernels based on four-step filtering criteria: 1) it should contain the predictive model (some kernels focus on data exploration only), 2) protected attribute is included in the training data, 3) at least five up-votes, and 4) ranked by up-votes. We used Kaggle API to collect these models and pass them through the filtering criteria. Finally, we select the top 6 models for each ensemble classifier from each dataset. In total, we created a benchmark of 168 ensembles across four datasets. We could not find certain classifiers on Kaggle for datasets like German Credit. To handle those, we use default models from scikit-learn to ensure we can evaluate across different datasets. The number of models mined is shown in Table I. Next, we present an overview of the datasets used in our benchmark. 

**Adult Census.** The dataset contains income and personal information about individuals [38]. We used sex as the protected attribute and male as the privileged class. The classification task predicts if a person makes over $50,000 in annual income.

**Titanic.** The dataset contains passenger data, such as gender, cabin class, etc., and is pre-split into train & test sets; however, the test set does not contain any instance of a male passenger surviving [39]. Hence, we only use the training set, with gender as the protected attribute and Female as the privileged class. The prediction task is whether a passenger survives.

**Bank Marketing.** The dataset contains bank customers’ personal information such as age, job type, etc. [40]. The protected attribute is age, where age > 25 is considered privileged class and age < 25 as unprivileged [13]. The prediction task determines whether a client will subscribe to a term deposit.

**German Credit.** This dataset contains personal and financial information about individuals who apply for loans at a bank [41]. We used the processed dataset [42] since most models in our benchmark used it. This version has nine attributes, such as sex, credit amount, etc. We choose sex as the protected attribute and male as the privileged class. The prediction task is whether an individual is a credit risk.

B. Measures

Multiple quantitative fairness and accuracy measures are available to evaluate a model. We use measures that have been previously used in literature [11], [13]. Let \( D = (T, S, Y) \) be a dataset where \( T \) is the training set, \( S \) is the protected attribute \( (S = 1, \text{if privileged group (p)}, \text{else } S = 0 \text{ (up)}) \) and \( Y \) is the classification label \( (Y = 1 \text{ if favorable label, else } Y = 0) \). Let \( \hat{Y} \) denote the prediction of an ML model. Next, we define our measures in terms of these notations.

1) **Accuracy Metrics:** We evaluate the performance of the models using accuracy and F1 metrics as defined below:

\[
\text{Accuracy} = \frac{\text{true positive} + \text{true negative}}{\text{total}}
\]

\[
F1 = 2 \times (\text{precision} \times \text{recall}) / (\text{precision} + \text{recall})
\]

where recall: \( TP/(TP + FN) \), precision: \( TP/(TP + FP) \)

2) **Fairness Measures:** Broadly, fairness metrics are divided into three categories [43]. We have selected a subset of these metrics representing the three categories without being exhaustive. Furthermore, we have followed the recommendations of Friedler et al. [13] in terms of metrics selection.

**Group fairness metrics:** Group fairness means similar predictive outcomes for protected attributes, e.g., race (Asian/White) on a group level.
TABLE I: Ensemble types used in our experiments

<table>
<thead>
<tr>
<th>Categories</th>
<th>Ensemble Types</th>
<th>Algorithms</th>
<th>Composition</th>
<th>Classifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential</td>
<td>Boosting</td>
<td>Construct n homogeneous estimators sequentially that improve predictions</td>
<td>Homogeneous</td>
<td>XGBoost</td>
</tr>
<tr>
<td></td>
<td></td>
<td>based on the previous estimator’s incorrect predictions</td>
<td>Homogeneous</td>
<td>AdaBoost</td>
</tr>
<tr>
<td></td>
<td>Bagging</td>
<td>Construct n parallel homogeneous models that are aggregated using</td>
<td>Homogeneous</td>
<td>Gradient Boosting</td>
</tr>
<tr>
<td></td>
<td>Voting</td>
<td>averaging</td>
<td>Homogeneous</td>
<td>Random Forest</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Homogeneous</td>
<td>ExtraTrees</td>
</tr>
<tr>
<td></td>
<td>Stacking</td>
<td>Construct a list of n heterogeneous user-specified weighted classifiers</td>
<td>Heterogeneous</td>
<td>Bagging Classifier</td>
</tr>
<tr>
<td></td>
<td></td>
<td>that are aggregated using majority voting or argmax</td>
<td>Voting Classifier</td>
<td></td>
</tr>
</tbody>
</table>

TABLE II: Summary of the datasets and the number of models collected for each in the benchmark

<table>
<thead>
<tr>
<th>Datasets</th>
<th>PA</th>
<th>Size</th>
<th># XGB</th>
<th># ADB</th>
<th># GBC</th>
<th># RF</th>
<th># ET</th>
<th># STK</th>
<th># VT</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult Census</td>
<td>sex</td>
<td>32681</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>33</td>
</tr>
<tr>
<td>Titanic ML</td>
<td>sex</td>
<td>891</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>42</td>
</tr>
<tr>
<td>Bank Marketing</td>
<td>age</td>
<td>41118</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>29</td>
</tr>
<tr>
<td>German Credit</td>
<td>sex</td>
<td>1000</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>17</td>
</tr>
</tbody>
</table>


Equal Opportunity Difference (EOD): This is defined as the difference of true-positive rates (TPR) between privileged and unprivileged groups [44].

\[ EOD = TPR_{up} - TPR_{p} \]

where TPR: \( TP/(TP + FN) \), FPR: \( FP/(FP + TN) \)

Average Odds Difference (AOD): This is defined as the mean of false-positive rate (FPR) difference and true-positive rate difference between unprivileged and privileged groups [27].

\[ AOD = \frac{1}{|A|} \sum_{a \in A} (\text{AOD}_a) = \frac{1}{|A|} \sum_{a \in A} \left( \frac{FP_{up} - FPR_{p}}{TPR_{up} - TPR_{p}} \right) \]

Disparate Impact (DI): This is defined as the ratio of the probability of unprivileged group vs. privileged group getting a favorable prediction [7].

\[ DI = P[\hat{\delta} = 1|Y = 0]/P[\hat{\delta} = 1|Y = 1] \]

We convert Disparate Impact (DI) to log scale to improve readability compared with other metrics.

Statistical Parity Difference (SPD): This is defined similar to DI but uses the difference between the probabilities. [45].

\[ SPD = P[\hat{\delta} = 1|Y = 0] - P[\hat{\delta} = 1|Y = 1] \]

Individual fairness metrics:

Theil Index (TI): It measures both the group and individual fairness [6]. It is defined using the following equation:

\[ TI = \frac{1}{n} \sum_{i=1}^{n} b_i \cdot \ln \frac{b_i}{\bar{b}_i} \text{, where } b_i = \hat{\delta}_i - \bar{\delta}_i + 1. \]

C. Experiment Design & Setup

Each ensemble model has specific requirements for training (e.g., XGBoost can handle Null values, but Random Forest cannot etc.) that we need to handle before we can evaluate them. We used the same preprocessing steps across all the kernels and datasets to ensure consistent comparison. Next, we evaluated the accuracy and fairness of base learners and the final ensemble level and analyzed the results.

For our data preprocessing, we start off by converting all non-numerical features to categorical data i.e. Binary or Ordinal (e.g., male: 1, female: 0) or non-binary levels, like Marital Status to Divorced: 0, Married: 1, Single: 2 etc.). Next, we remove missing values from our datasets and convert continuous sensitive attributes to categorical (e.g., age > 25: 1, age < 25: 0 corresponding to old and young, respectively). These preprocessing steps are necessary for most ensembles and the AIF360 toolkit. We denote the privileged and unprivileged groups and the favorable label for each dataset separately. For example, in Titanic dataset, male is the unprivileged group and the favorable label is Survived: 1 i.e. the individual survived the Titanic crash. The groups and the labels have been chosen as seen before in literature [13], [46]. Finally, the dataset is shuffled and split into train and test sets using a 70% – 30% split. For each dataset, we have selected the top 6 kernels by upvotes. We run the preprocessing steps discussed before training the model to evaluate based on accuracy and fairness metrics. We use five fairness metrics and two accuracy measures to generate results for each model. These experiments are repeated ten times, and the mean is reported [11]. We used the IBM AIF 360 Fairness Toolkit to calculate the fairness metrics. Finally, a non-zero value for fairness metrics suggests a bias in the model. A positive value of a fairness metric suggests the model bias is against the privileged group and vice-versa.

IV. FAIRNESS IN ENSEMBLES AND ITS COMPOSITION

In this section, we explore the state of fairness in ensembles and its composition in all popular ensemble methods.

A. State of fairness in ensemble models

Before understanding the composition of fairness in ensembles, we first investigate how different ensemble techniques impact fairness (RQ1). Are certain ensemble classifiers more unfair? Does the architecture of an ensemble method (stacking, boosting, etc.) contribute to fairness? Does any particular ensemble classifier exhibit a better fairness-accuracy trade-off? To answer these questions, we experiment to evaluate the fairness of ensemble models using a diverse set of metrics. Table [III] shows the mean fairness for all ensembles. Figure [1] illustrates the cumulative fairness for all the 168 models.
We found that ensemble models show similar fairness patterns specific to each dataset and that some ensemble classifiers exhibit more unfairness than others. To investigate whether these differences are significant and directly impacted by ensemble type, we compare their fairness and accuracy using Scott-Knott ranking test. The Scott-Knott test compares two distributions using a recursive bi-clustering algorithm and provides a rank; if the difference between them is insignificant, they get the same rank. Otherwise, a higher rank denotes better performance. In our experiment, the Scott-Knott test ranks the classifiers from 1st to 5th (some classifiers are assigned the same rank) for each metric. Even though each ensemble classifier’s unfairness pattern is similar to the dataset, certain ensembles are fairer than others.

Finding 1: Among all the ensemble models, XGBoost exhibits the best accuracy-fairness trade-off.

From Table III, we can see that the fairness of the ensemble classifiers varies across the datasets. Moreover, we see that the fairness of the classifiers within an ensemble type also varies. For instance, AdaBoost is the most unfair model in the Adult dataset, while XGBoost is the fairest, both belonging to boosting. However, we observe that for the more biased Titanic and Adult datasets, XGBoost has the top rank in 8 out of 10 fairness metrics with a mean fairness rank of 1.4 and 1.6, respectively. Yet, XGBoost is ranked 2nd and 1st in terms of accuracy, respectively. The average performance change in accuracy and f1 score is less than 0.01. On the contrary, the cumulative mean fairness for XGBoost models is 14% more than the next most fair model (GBC) in Titanic. For the other datasets, we observe a similar pattern; however, the difference is lesser. Upon further investigation, we found that boosting method and base learner design are responsible for the fairer performance of XGBoost. Homogeneous ensembles use decision trees as the base learner, and the construction of these trees differs among them. For example, the depth of the decision tree in AdaBoost, GradientBoosting, XGBoost is one, three and six respectively. Lower tree depth means fewer features are selected, which has been shown to often increase unfairness [13], [48]. This is also one of the reasons behind AdaBoost being the most unfair model (the lowest rank in 13 out of 16 metrics). Importantly, we found that the fairness of an ensemble is determined by the composition of fairness within these base learners and the learning method (boosting, bagging, etc.). In the next section, we investigate the properties of the base learners to understand how to compose fair ensembles in more detail.

Finding 2: Fairness measures show more instability compared to accuracy metrics.

---

**TABLE III:** Fairness and accuracy comparison of all ensemble ML classifiers across the datasets in our benchmark. The ranks were calculated using the Scott-Knott test. Each cell depicts the median score. Darker, lighter, light, lightest and white colored cell denotes the first, the second, the third, fourth, and lowest rank, respectively. The rank ranges from 1 to 5.
Prior works have shown that ensembles improve the stability of accuracy metrics by aggregating multiple learners trained on subsets of data (bootstrap sampling) [44], [49]. Intuitively, we explore the variance exhibited by fairness metrics in ensembles. Interestingly, despite aggregating multiple learners, the stability of fairness measures in ensembles still suffers, especially in smaller datasets. This is attributed to the change in data distribution after random train/test splits in smaller datasets [11], [13]. For larger datasets (Adult, Bank Marketing), the standard deviation for all fairness measures is less than 0.02. For smaller datasets, the average standard deviation of the metrics is shown in Figure 2. Firstly, we observe that the stability of fairness metrics remains consistent between all the ensembles for a specific dataset. Furthermore, we observed that group fairness measures exhibit higher variability than individual measures (TI). Surprisingly, heterogeneous models also exhibit instability despite using dissimilar learners to reduce variance. From Figure 3, we also see that the volatility in fairness metrics is greater than in accuracy metrics for homogeneous models. Given a random train/test split, it might cause the model to seem fairer than it is. Hence, even with improved stability in fairness compared to non-ensemble methods, developers should evaluate the training set and repeat training over multiple runs in ensembles.

**Finding 3:** Libraries do not provide API support to measure fairness of base learners in ensembles

Biswas and Rajan [13] discussed that hyperparameter optimization goals induce unfairness. Whereas ensemble models also have parameters that can be tuned, heterogeneous models have a special case where the developer needs to select the individual base learners. Libraries do not provide any recommendations to developers, who try to select a diverse set of learners to improve accuracy. However, this might not always result in a fair ensemble. For example, BM-STK3 has Statistical Parity Difference (SPD) of 0.13 but if we drop GuassianNB learner, the SPD improves to 0.11 with better accuracy. Heterogeneous Voting models uses weighted voting, while Stacking uses a meta-learner to select the best configuration of learners for each prediction. Since libraries do not provide API support to measure the fairness of base learners, especially in combinations with other learners at the prediction level, it provides developers with little information on how to weigh and select individual learners to generate the final predictions. Similarly, understanding fairness composition in base learners of homogeneous models can help the developer identify fairness issues such as bias in specific features (e.g., decision trees learners in random forest randomly select features). Therefore, API support to measure fairness in base learners can help developers better understand & detect unfairness in ensembles.

**B. How does fairness compose in ensembles?**

In this section, we investigate the composition of fairness in ensembles. We postulate that the underlying unfairness of ensembles is a product of the composition of fairness in base learners and the learning method. All homogeneous models use a decision tree as the base learner, whereas heterogeneous models can take any ML model. We investigate how fairness composes in these base learners and how it is propagated by the learning techniques (RQ2). In general, our findings show that the complexity of base learners significantly impacts the fairness of ensembles and that more research is needed to develop fair learning techniques in ensembles.

**Finding 4:** The unfairness of homogeneous ensembles is caused by the complexity of the base learner and dataset characteristics.
In Figure 3 we plotted the most biased and the least biased homogeneous ensembles in our benchmark. We see how the fairness of the base learners directly contributes to the fairness of ensembles during training. However, we observe different fairness composition patterns in terms of the datasets. For example, in Titanic, the fairness of the boosting ensembles improves while the opposite is observed for some fairness metrics in German Credit. The variation in fairness patterns is also seen in specific classifiers. From Figure 1 we observe that all TM-XGB models show similar bias except, TM-XGB4. We investigate the difference in unfairness by comparing all the parameters of the base learner decision tree with the other XGB models and found that TM-XGB4 uses a shallow decision tree with max_depth : 2, which is causing the unfairness to amplify. The model construct is shown below:

```python
model = XGBClassifier(n_estimators= 500, max_depth=2, subsample =0.5, learning_rate=0.1)
```

MaxDepth sets the maximum depth of the decision tree. The depth of the tree is defined as the number of splits (nodes), where the feature to be split is chosen based on the highest information gain among the features. Deeper trees are more complex and reduce errors [50]. For XGBoost models, the default depth is 6. Our analysis showed that the protected attribute (Sex) has the highest information gain among all the features in the Titanic dataset. Therefore, the protected attribute is the most important feature to split on at the tree’s root, resulting in a high degree of unfairness in TM-XGB4. Base learners in all boosting models use the best feature to split, which improves accuracy. However, it has been shown that unfairness is encoded in specific features [51]. If these features are also among the best features of a dataset, a shallower ensemble will be more unfair due to a reduced number of features. This corroborates similar observations in the literature [12], [48]. We observe the same pattern for all boosting models. For example, AdaBoost and GradientBoosting exhibit more bias than XGBoost because of shallower base learners (1 and 3, respectively). In Figure 1 TM-GBC1 is fairer because the learner is deeper (depth:5).

Finally, further analysis of the properties of a decision tree suggests that regularization parameters like min samples leaf and max-leaf nodes also impact tree depth, hence affecting the fairness of the ensemble. Therefore, it is important to carefully balance the complexity of the tree-based base learners for homogeneous models with the fairness outcomes, especially with the underlying properties of the data.

### Finding 5: Gradient-based composition propagates more unfairness compared to Adaptive boosting models.

Fig. 4: Cumulative fairness in adaptive vs gradient learning

We have established that base learners and the underlying data properties influence the unfairness of homogeneous models. However, boosting models additionally use a learning technique to sequentially improve the predictions of the model. XGBoost and GradientBoosting models uses Gradient-based optimization and AdaBoost uses an adaptive weighting technique. We compare these techniques by training the boosting models on the same base learner decision tree (depth:6). We only use XGBoost and choose this depth in our experiment since our analysis (Table III) showed that it is the most fair boosting model. The results are shown in Figure 4. For all the models, except GC1, we see that Adaptive learning is fairer than Gradient optimization. We use the Scott-Knott rank test to test statistical significance. Accordingly, we observed that adaptive learning outperformed gradient optimization in all datasets, except German Credit where the difference was not statistically significant. Consequently, we can see that adaptive learning propagates less bias in highly biased datasets. Our analysis should help guide further research into designing fair learning techniques for boosting ensembles.
V. FAIR ENSEMBLES DESIGN

In §IV we found that base learners of ensembles propagate bias. Many bias mitigation techniques applied during model training (inprocess) have been successful [24], [25], [32]. Techniques applied during the model training phase can assist developers in improving the fairness of ML software. These works have also established the role of hyperparameters in mitigating and amplifying fairness bugs (unfairness) in ML software. If we understand and identify what ensemble parameters and design choices affect the fairness, we can mitigate inherent bias in ensembles. Moreover, it will help developers, and libraries better explain fairness bugs in the ensemble hyperparameter space. This section explores the hyperparameter design space for ensembles to boost fairness performance. We have found that some hyperparameters directly affect the fairness of ensembles. Specifically, we evaluate how ensembles can be designed to be fair using ensemble hyperparameters summarized in Table IV. We use the Scott-Knott test to determine the significance of our results. Our findings provide a comprehensive review of all ensemble hyperparameters.

Finding 6: Developers should carefully choose dropout regularization to balance fairness and overfitting.

Our analysis shows that dropout impacts fairness in relation to the underlying data properties. Vinayak and Gilad-Bachrach [52] proposed DART, a dropout technique derived from deep neural networks, for boosted trees. An ensemble of boosted regression trees suffers from over-specification, i.e., the trees added at the end have little contribution to the final result [52]. DART alleviates this by constructing the next tree from the residuals of a random sample of the previous trees. In XGBoost, the rate-drop ([0-1]) parameter controls this sampling rate. No trees are dropped on the lower end of this rate, while on the higher extreme, all trees are dropped. We investigate the efficacy of DART with \( \text{rate-drop} = 0.5 \), in reducing unfairness in boosting models by comparing it with the default XGBoost booster gbtree. We analyze the change in performance and fairness of dropout in Figure 5.

From figure 5, we can see that dropout can impact fairness of boosting models. For example, in Adult dataset, initial trees exhibit less unfairness than using dropout. Using dropout, the subsequent trees only learn on a random sample of initial trees, which in this case are fairer. This improves the fairness of the models. The opposite is observed in Titanic dataset. In both scenarios, the change in accuracy is less than 0.1, but a significant impact is seen on fairness. Therefore, developers should be cautious about the effect of regularization on fairness. More research is needed to understand fairness-overfitting trade-off, and develop fair regularization methods.

Finding 7: Randomness in feature splitting does not improve fairness in bagging models.

Random Forest improves the variance of the model by introducing randomness to the process of model building by randomly selecting features. Extra Trees introduces additional randomness by randomly finding the splits for each feature and then selecting the best split from them, i.e., independent of the target variable. In contrast, Random Forest finds the best split for each feature which has been shown to improve accuracy [53]. However, no work has studied its effect on fairness. Here, we ask whether randomness at the feature splitting level causes bagging models to be unfair.

To investigate this, we compare Random Forest and Extra Tree models in our benchmark. We keep the rest of the parameters and data split the same. Each model is run ten times, and the mean is reported in Table V. For all datasets except German Credit, the test showed that Extra Tree models with random splits were more biased compared to optimal splits (RF). This is a key finding because this suggests that a split chosen independently of the target is still more unfair than an optimal split. However, in the fairer dataset (German Credit), we observe no difference in fairness. Regarding bias mitigation methods, our results suggest randomness in feature split-point might not be an effective way to tackle bias in decision tree-based models.

Finding 8: The uncertainty in classifiers can have a large impact on fairness in voting classifiers.

A Voting classifier is an ensemble method where the prediction is based on the probabilities of each base learner within the ensemble. Voting classifiers are of two types, Soft and Hard Voting. In hard Voting, the label with the majority of votes from the base learners is the final prediction, whereas, in soft voting, it is based on the average of the probabilities of each output class. If the average probability of a class is less than 0.5, class 0 is predicted and vice-versa. We investigate the effect of the voting type on fairness and found that the
uncertainty in the model prediction can have a large impact on fairness. For instance, AC-VT5 uses soft voting with Logistic Regression (LR), Random Forest (RF), KNN, and Decision Tree (DT) as base classifiers. As shown in Table VI, DT introduces significant unfairness when used in soft voting compared to hard. We found that DT has an output class probability of either 1 or 0 while other classifiers are in the range [0,1]. Figure 6 shows the output probabilities for AC-VT5. We observe that DT has higher extreme probabilities compared to others. In this case, the average is skewed by the extremely probabilities of DT. This changes the prediction for 558 out of 9049 test samples. And since DT is less fair than other classifiers, the overall unfairness also increases. For hard voting, equal weight is given to each classifier. In that case, the other three classifiers, which are fairer, won the majority vote. For some models, soft voting was fairer than hard, e.g., AC-VT3, which shows that base learners’ uncertainties can impact fairness in both voting types. This suggests the need to develop frameworks to measure model uncertainties and their fairness at a component level to aid developers in designing fair voting ensembles. Our analysis should also encourage further research in fairness-aware weighting techniques to handle fairness issues arising from model uncertainties.

Finding 9: Two-layer stacking can significantly reduce unfairness.

All of theTitanic ML stacking models shown in Figure 1 exhibit similar bias except TM-STK5, which is the least biased model for all fairness metrics except Thiel Index (TI). On closer inspection, we found out that TM-STK5 uses a two-layered stacking approach where a second layer of base learners act as the meta-learner, which causes the model to be fairer. The model construct is shown below:

```r
layer_one_estimators = [("rf_1"), RandomForestClassifier( n_estimators=40, random_state=42)], ("knn_1", KNeighborsClassifier(n_neighbors=6))
layer_two_estimators = [("rf_2"), RandomForestClassifier( n_estimators=40, random_state=42)], ("xg_2", XGBClassifier( objective = 'binary:logistic', colsample_bytree = 0.8, learning_rate = 0.3, max_depth = 7, min_child_weight = 3, n_estimators = 100, subsample = 0.6))
model = StackingClassifier( estimators=layer_two_estimators, final_estimator=XGBClassifier(n_estimators = 100))
```

We validate our finding by training all stacking models in our benchmark using this two-layered nested stacking approach. To ensure consistency, we did not change the kernel’s feature set or any preprocessing method. The results are shown in Figure 7. For all stacking models in our benchmark, every model significantly improved in all fairness measures except Thuil Index. Achieving fairness in terms of all fairness metrics is difficult [8, 54]. Moreover, accuracy measures did not degrade significantly. For example, TM-STK6 improved SPD scores by 28% while accuracy dropping only 4.68%. Overall across all datasets, the accuracy score dropped by 6.50% while...
In stacking, the first layer uses a list of base learners to generate a set of predictions and a meta-learner to learn from them. However, instead of using a single classifier as the meta-learner, the predictions are fed into another set of base learners in a two-layered approach. This ensures that the outcome is not based on a single meta-learner. Instead, the second layer of learners creates a new set of predictions, which are then fed into the second layer’s final meta-classifier. We compare this approach to simply using an ensemble model (fed into the second layer’s final meta-classifier. We compare this approach to stacking (Top) and XGB as meta-learner (Bottom). A positive change indicates performance/fairness increase and vice-versa.

In stacking, the SPD improved by 31.8%.

In this study, we undertook an important task of understanding the composition of fairness in ensemble learning that has gained popularity recently [12], [13]. However, most works focus on fairness mitigation in single classifiers, which fail to capture inherent bias in ensembles. Our work has shown that fair ensembles can be designed without mitigation techniques. Our research also identifies unfairness causes in different ensembles and their interplay with the input space, which would guide fairness bug localization and repair in ensemble learning. For example, we reported fairness patterns in individual learners that can induce bias in ensembles such as tree depth, min leaf node samples, etc. These can also be leveraged for fairness-improving interventions such as feature selection, data preprocessing, etc. Overall, our result would draw attention to the fairness of ensembles which are popular learning algorithms but mostly overlooked by the community.

Moreover, research in SE showed the impact of hyperparameters on fairness, and their role in helping developers mitigate bias during model training [12], [25], [32]. We extend that to explore the hyperparameter space for ensembles to guide developers to design fair ensembles using currently available composition and configuration. We discussed the effect of regularization on fairness which should encourage the development of fairness-aware regularization techniques and investigation of the trade-off between fairness mitigation and overfitting, similar to the accuracy-fairness trade-off. We found that many ensemble models do not have library support to monitor the fairness of individual learners. Our work would encourage the development of tools and API support to improve the transparency of ML software to address fairness concerns.

VII. Threats to Validity

Benchmark: We ensure quality of the benchmark by collecting only high-quality kernels from Kaggle (atleast 5 upvotes). Additionally, we only consider models that are runnable, include the protected attribute in training, and have an accuracy greater than 65%, similar to [13]. Finally, we select the top 6 (upvotes) models for each ensemble type.

Sampling Bias: To the best of our knowledge, this is the most extensive review of popular ensembles. Moreover, conclusions are supported by statistical tests across four datasets. However, they may change slightly if other datasets are used.

External validity: To avoid the threat of non-generalized findings, we conduct experiments on four different datasets for each ensemble type and compare across multiple ensemble algorithms for both boosting and bagging. Moreover, we use multiple fairness metrics and verify our results by running the experiment multiple (ten) times and using mean of the values.

VIII. Related Work

a) Fairness in ML classification: The ML community has proposed multiple methods to measure [2], [4], [5], [8], [27], [55] and mitigate unfairness in ML models [3], [8], [9], [24], [56]. However, most of these works have focused on the theoretical evaluation of fairness. Recently, the SE community has increasingly shown interest in fairness in ML software [10]. Empirical studies have investigated the characteristics of biased models, unfairness in ML pipelines, compared mitigation strategies and developer concerns about fairness [11]–[13], [57]. Some research in SE has focused on fairness testing to generate test data inputs and uncover fairness violations [1], [14], [58], [59]. Finally, a body of work has identified unfairness in data and proposed appropriate mitigation techniques [25], [37], [60], [61].

b) Ensemble Fairness: Grgic-Hlaca et al. [31] investigated the impact of fairness in random-selection based ensemble. They showed theoretically that its fairness at the ensemble level is always fairer than its components. Wang et al. [62] studied the composition of fairness in multi-component recommender systems and presented conditions under which individual components compose fairness. AdaFair [30] proposed a fairness-aware AdaBoost model where unfairly classified instances were up-weighted. A recent work
analyzed and compared seven ML models to show that ensembles were fairer than individual classifiers. Feffer et al. conducted an empirical study to analyze modular ensembles. They developed a library to find the best configuration using any combination of ensembles and mitigators. In Fair Pipelines [15], the authors explored the propagation of fairness in multi-stage pipelines where a set of decisions impacts the final result, e.g., the hiring process. MAAT [63] proposes an ensemble approach to improve fairness performance by separately combining models optimized for fairness and accuracy. Finally, Tizpaz-Niari et al. studied the parameter space of ML algorithms and its impact on fairness [32]. This work is the closest to our study; however, it proposed a testing approach to tune the parameters for achieving fairness and did not consider ensembles (except random forest). Our work has focused on comprehensively evaluating fairness composition in all popular ensemble models and how the different algorithmic design configurations (parameters) impact fairness.

IX. CONCLUSION
Ensembles are widely used for predictive tasks due to superior performance. However, most approaches to measuring fairness and mitigation focus on single classifiers. In this paper, we conduct an empirical study to evaluate the composition of fairness in popular ensemble techniques. The results showed that base learners induce bias in ensembles and that we can mitigate inherent bias in ensembles by using certain base learner configurations and appropriate parameters. Lastly, works have shown the need to support developers during model training in mitigating bias. Our analysis of the hyper-parameter space should help developers build fairness-aware ensembles and automated tools to detect bias in ensembles.

REFERENCES


