

Certifying the Fairness of KNN in the Presence of Dataset Bias^{*}

Yannan Li ✉, Jingbo Wang, and Chao Wang

University of Southern California, Los Angeles CA 90089, USA
{yannanli, jingbow, wang626}@usc.edu

Abstract. We propose a method for certifying the fairness of the classification result of a widely used supervised learning algorithm, the k -nearest neighbors (KNN), under the assumption that the training data may have historical bias caused by systematic mislabeling of samples from a protected minority group. To the best of our knowledge, this is the first certification method for KNN based on three variants of the fairness definition: individual fairness, ϵ -fairness, and label-flipping fairness. We first define the fairness certification problem for KNN and then propose sound approximations of the complex arithmetic computations used in the state-of-the-art KNN algorithm. This is meant to lift the computation results from the concrete domain to an abstract domain, to reduce the computational cost. We show effectiveness of this *abstract interpretation* based technique through experimental evaluation on six datasets widely used in the fairness research literature. We also show that the method is accurate enough to obtain fairness certifications for a large number of test inputs, despite the presence of historical bias in the datasets.

1 Introduction

Certifying the fairness of the classification output of a machine learning model has become an important problem. This is in part due to a growing interest in using machine learning techniques to make socially sensitive decisions in areas such as education, healthcare, finance, and criminal justice systems. One reason why the classification output may be biased against an individual from a protected minority group is because the dataset used to train the model may have historical bias; that is, there is systematic mislabeling of samples from the protected minority group. Thus, we must be extremely careful while considering the possibility of using the classification output of a machine learning model, to avoid perpetuating or even amplifying historical bias.

One solution to this problem is to have the ability to certify, with certainty, that the classification output $y = M(x)$ for an individual input x is fair, despite that the model M is learned from a dataset T with historical bias. This is a

^{*} This work was partially funded by the U.S. National Science Foundation grants CNS-1702824, CNS-1813117 and CCF-2220345.

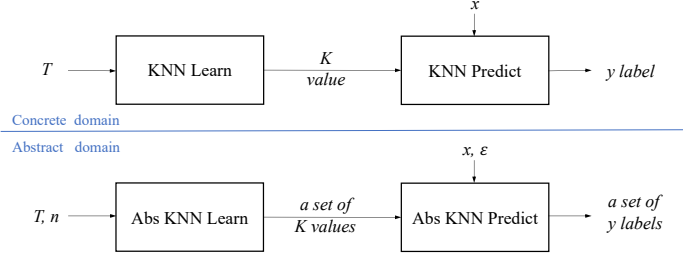


Fig. 1. FAIRKNN: our method for certifying fairness of KNNs with label bias.

form of *individual fairness* that has been studied in the fairness literature [14]; it requires that the classification output remains the same for input x even if historical bias were not in the training dataset T . However, this is a challenging problem and, to the best of our knowledge, techniques for solving it efficiently are still severely lacking. Our work aims to fill the gap.

Specifically, we are concerned with three variants of the fairness definition. Let the input $x = \langle x_1, \dots, x_D \rangle$ be a D -dimensional input vector, and $\mathcal{P}$ be the subset of vector indices corresponding to the *protected* attributes (e.g., race, gender, etc.). The first variant of the fairness definition is *individual fairness*, which requires that similar individuals are treated similarly by the machine learning model. For example, if two individual inputs x and x' differ only in some protected attribute x_i , where $i \in \mathcal{P}$, but agree on all the other attributes, the classification output must be the same. The second variant is ϵ -*fairness*, which extends the notion of individual fairness to include inputs whose un-protected attributes differ and yet the difference is bounded by a small constant (ϵ). In other words, if two individual inputs are almost the same in all unprotected attributes, they should also have the same classification output. The third variant is *label-flipping fairness*, which requires the aforementioned fairness requirements to be satisfied even if a biased dataset T has been used to train the model in the first place. That is, as long as the number of mislabeled elements in T is bounded by n , the classification output must be the same.

We want to certify the fairness of the classification output for a popular supervised learning technique called the k -nearest neighbors (KNN) algorithm. Our interest in KNN comes from the fact that, unlike many other machine learning techniques, KNN is a *model-less* technique and thus does not have the high cost associated with training the model. Because of this reason, KNN has been widely adopted in real-world applications [1, 4, 16, 18, 23, 29, 36, 45, 46]. However, obtaining a fairness certification for KNN is still challenging and, in practice, the most straightforward approach of *enumerating all possible scenarios* and then *checking if the classification outputs obtained in these scenarios agree* would have been prohibitively expensive.

To overcome the challenge, we propose an efficient method based on the idea of *abstract interpretation* [10]. Our method relies on sound approximations to

analyze the arithmetic computations used by the state-of-the-art KNN algorithm both accurately and efficiently. Figure 1 shows an overview of our method in the lower half of this figure, which conducts the analysis in an abstract domain, and the default KNN algorithm in the upper half, which operates in the concrete domain. The main difference is that, by staying in the abstract domain, our method is able to analyze a large set of possible training datasets (derived from T due to n label-flips) and a potentially-infinite set of inputs (derived from x due to ϵ perturbation) symbolically, as opposed to analyze a single training dataset and a single input concretely.

To the best of our knowledge, this is the first method for KNN fairness certification in the presence of dataset bias. While Meyer et al. [26, 27] and Drews et al. [12] have investigated robustness certification techniques, their methods target decision trees and linear regression, which are different types of machine learning models from KNN. Our method also differs from the KNN data-poisoning robustness verification techniques developed by Jia et al. [20] and Li et al. [24], which do not focus on *fairness* at all; for example, they do not distinguish *protected* attributes from *unprotected* attributes. Furthermore, Jia et al. [20] consider the prediction step only while ignoring the learning step, and Li et al. [24] do not consider label flipping. Our method, in contrast, considers all of these cases.

We have implemented our method and demonstrated the effectiveness through experimental evaluation. We used all of the six popular datasets in the fairness research literature as benchmarks. Our evaluation results show that the proposed method is efficient in analyzing complex arithmetic computations used in the state-of-the-art KNN algorithm, and is accurate enough to obtain fairness certifications for a large number of test inputs. To better understand the impact of historical bias, we also compared the fairness certification success rates across different demographic groups.

To summarize, this paper makes the following contributions:

- We propose an abstract interpretation based method for efficiently certifying the fairness of KNN classification results in the presence of dataset bias. The method relies on sound approximations to speed up the analysis of both the learning and the prediction steps of the state-of-the-art KNN algorithm, and is able to handle three variants of the fairness definition.
- We implement the method and evaluate it on six datasets that are widely used in the fairness literature, to demonstrate the efficiency of our approximation techniques as well as the effectiveness of our method in obtaining sound fairness certifications for a large number of test inputs.

The remainder of this paper is organized as follows. We first present the technical background in Section 2 and then give an overview of our method in Section 3. Next, we present our detailed algorithms for certifying the KNN prediction step in Section 4 and certifying the KNN learning step in Section 5. This is followed by our experimental results in Section 6. We review the related work in Section 7 and, finally, give our conclusion in Section 8.

2 Background

Let L be a supervised learning algorithm that takes the training dataset T as input and returns a learned model $M = L(T)$ as output. The training set $T = \{(x, y)\}$ is a set of labeled samples, where each $x \in \mathcal{X} \subseteq \mathbb{R}^D$ has D real-valued attributes, and the $y \in \mathcal{Y} \subseteq \mathbb{N}$ is a class label. The learned model $M : \mathcal{X} \rightarrow \mathcal{Y}$ is a function that returns the classification output $y' \in \mathcal{Y}$ for any input $x' \in \mathcal{X}$.

2.1 Fairness of the Learned Model

We are concerned with *fairness* of the classification output $M(x)$ for an individual input x . Let $\mathcal{P}$ be the set of vector indices corresponding to the protected attributes in $x \in \mathcal{X}$. We say that x_i is a protected attribute (e.g., race, gender, etc.) if and only if $i \in \mathcal{P}$.

Definition 1 (Individual Fairness). *For an input x , the classification output $M(x)$ is fair if, for any input x' such that (1) $x_j \neq x'_j$ for some $j \in \mathcal{P}$ and (2) $x_i = x'_i$ for all $i \notin \mathcal{P}$, we have $M(x) = M(x')$.*

It means two individuals (x and x') differing only in some protected attribute (e.g., gender) but agreeing on all other attributes must be treated equally. While being intuitive and useful, this notion of fairness may be too narrow. For example, if two individuals differ in some unprotected attributes and yet the difference is considered *immaterial*, they must still be treated equally. This can be captured by ϵ -fairness.

Definition 2 (ϵ -Fairness). *For an input x , the classification output $M(x)$ is fair if, for any input x' such that (1) $x_j \neq x'_j$ for some $j \in \mathcal{P}$ and (2) $|x_i - x'_i| \leq \epsilon$ for all $i \notin \mathcal{P}$, we have $M(x) = M(x')$.*

In this case, such inputs x' form a set. Let $\Delta^\epsilon(x)$ be the set of all inputs x' considered in the ϵ -fairness definition. That is, $\Delta^\epsilon(x) := \{x' \mid x_j \neq x'_j \text{ for some } j \in \mathcal{P}, |x_i - x'_i| \leq \epsilon \text{ for all } i \notin \mathcal{P}\}$. By requiring $M(x) = M(x')$ for all $x' \in \Delta^\epsilon(x)$, ϵ -fairness guarantees that a larger set of individuals similar to x are treated equally.

Individual fairness can be viewed as a special case of ϵ -fairness, where $\epsilon = 0$. In contrast, when $\epsilon > 0$, the number of elements in $\Delta^\epsilon(x)$ is often large and sometimes infinite. Therefore, the most straightforward approach of certifying fairness by enumerating all possible elements in $\Delta^\epsilon(x)$ would not work. Instead, any practical solution would have to rely on abstraction.

2.2 Fairness in the Presence of Dataset Bias

Due to historical bias, the training dataset T may have contained samples whose output are unfairly labeled. Let the number of such samples be bounded by n .

We assume that there are no additional clues available to help identify the mis-labeled samples. Without knowing which these samples are, fairness certification must consider all of the possible scenarios. Each scenario corresponds to a de-biased dataset, T' , constructed by flipping back the incorrect labels in T . Let $\text{dBias}_n(T) = \{T'\}$ be the set of these possible de-biased (clean) datasets. Ideally, we want all of them to lead to the same classification output.

Definition 3 (Label-flipping Fairness). *For an input x , the classification output $M(x)$ is fair against label-flipping bias of at most n elements in the dataset T if, for all $T' \in \text{dBias}_n(T)$, we have $M'(x) = M(x)$ where $M' = L(T')$.*

Label-flipping fairness differs from and yet complements individual and ϵ -fairness in the following sense. While individual and ϵ -fairness guarantee equal output for similar inputs, label-flipping fairness guarantees equal output for similar datasets. Both aspects of fairness are practically important. By combining them, we are able to define the entire problem of certifying fairness in the presence of historical bias.

To understand the complexity of the fairness certification problem, we need to look at the size of the set $\text{dBias}_n(T)$, similar to how we have analyzed the size of $\Delta^\epsilon(x)$. While the size of $\text{dBias}_n(T)$ is always finite, it can be astronomically large in practice. Let q is the number of unique class labels and m be the actual number of flipped elements in T . Assuming that each flipped label may take any of the other $q - 1$ possible labels, the total number of possible *clean sets* is $\binom{|T|}{m} \cdot (q - 1)^m$ for each m . Since $m \leq n$, $|\text{dBias}_n(T)| = \sum_{m=1}^n \binom{|T|}{m} \cdot (q - 1)^m$. Again, the number of elements in $\text{dBias}_n(T)$ is too large to enumerate, which means any practical solution would have to rely on abstraction.

3 Overview of Our Method

Given the tuple $\langle T, \mathcal{P}, n, \epsilon, x \rangle$, where T is the training set, $\mathcal{P}$ represents the protected attributes, n bounds the number of biased elements in T , and ϵ bounds the perturbation of x , our method checks if the KNN classification output for x is fair.

3.1 The KNN Algorithm

Since our method relies on an *abstract interpretation* of the KNN algorithm, we first explain how the KNN algorithm operates in the concrete domain (this subsection), and then lift it to the abstract domain in the next subsection.

As shown in Fig. 2, KNN has a prediction step where `KNN_predict` computes the output label for an input x using T and a given parameter K , and a learning step where `KNN_learn` computes the K value from the training set T .

Unlike many other machine learning techniques, KNN does not have an explicit model M ; instead, M can be regarded as the combination of T and K .

```

1 func KNN_predict( $T, K, x$ ) {
2   Let  $T_x^K$  = the  $K$  nearest neighbors of  $x$  in  $T$ ;
3   Let  $Freq(T_x^K)$  = the most frequent label in  $T_x^K$ ;
4   return  $Freq(T_x^K)$ ;
5 }
6
7 func KNN_learn( $T$ ) {
8   for (each candidate  $k$  value) { // conducting  $p$ -fold cross validation
9     Let  $\{G_i\}$  = a partition of  $T$  into  $p$  groups of roughly equal size;
10    Let  $err_i^k = \{(x, y) \in G_i \mid y \neq \text{KNN\_predict}(T \setminus G_i, k, x)\}$  for each  $G_i$ ;
11  }
12  Let  $K = \arg \min_k \frac{1}{p} \sum_{i=1}^p \frac{|err_i^k|}{|G_i|}$ ;
13  return  $K$ ;
14 }

```

Fig. 2. The KNN algorithm, consisting of the prediction and learning steps.

Inside `KNN_predict`, the set T_x^K represents the K -nearest neighbors of x in the dataset T , where distance is measured by Euclidean (or Manhattan) distance in the input vector space. $Freq(T_x^K)$ is the most frequent label in T_x^K .

Inside `KNN_learn`, a technique called *p-fold cross validation* is used to select the optimal value for K , e.g., from a set of candidate k values in the range $[1, |T| \times (p-1)/p]$ by minimizing classification error, as shown in Line 12. This is accomplished by first partitioning T into p groups of roughly equal size (Line 9), and then computing err_i^k (a set of misclassified samples from G_i) by treating G_i as the evaluation set, and $T \setminus G_i$ as the training set. Here, an input $(x, y) \in G_i$ is “misclassified” if the expected output label, y , differs from the output of `KNN_predict` using the candidate k value.

3.2 Certifying the KNN Algorithm

Algorithm 1 shows the top-level procedure of our fairness certification method, which first executes the KNN algorithm in the concrete domain (Lines 1-2), to obtain the default K and y , and then starts our analysis in the abstract domain.

Algorithm 1: Our method for certifying fairness of KNN for input x .

```

1  $K = \text{KNN\_learn}(T)$ ;
2  $y = \text{KNN\_predict}(T, K, x)$ ;
3  $KSet = \text{abs\_KNN\_learn}(T, n)$ ;
4 for each  $K \in KSet$  do
5   if  $\text{abs\_KNN\_predict\_same}(T, n, K, x, y) = \text{False}$  then
6     return unknown;
7   end if
8 end for
9 return certified;

```

In the *abstract* learning step (Line 3), instead of considering T , our method considers the set of all clean datasets in $\text{dBias}_n(T)$ symbolically, to compute the set of possible optimal K values, denoted $KSet$.

In the *abstract* prediction step (Lines 4-8), for each K , instead of considering input x , our method considers all perturbed inputs in $\Delta^\epsilon(x)$ and all clean datasets in $\text{dBias}_n(T)$ symbolically, to check if the classification output always stays the same. Our method returns “certified” only when the classification output always stays the same (Line 9); otherwise, it returns “unknown” (Line 6).

We only perturb numerical attributes in the input x since perturbing categorical or binary attributes often does not make sense in practice.

In the next two sections, we present our detailed algorithms for abstracting the prediction step and the learning step, respectively.

4 Abstracting the KNN Prediction Step

We start with abstract KNN prediction, which is captured by the subroutine *abs_KNN_predict_same* used in Line 5 of Algorithm 1. It consists of two parts. The first part (to be presented in Section 4.1) computes a superset of T_x^K , denoted *overNN*, while considering the impact of ϵ perturbation of the input x . The second part (to be presented in Section 4.2) leverages *overNN* to decide if the classification output always stays the same, while considering the impact of label-flipping bias in the dataset T .

4.1 Finding the K -Nearest Neighbors

To compute *overNN*, which is a set of samples in T that *may be* the K nearest neighbors of the test input x , we must be able to compute the distance between x and each sample in T .

This is not a problem at all in the concrete domain, since the K nearest neighbors of x in T , denoted T_x^K , is fixed and is determined solely by the Euclidean distance between x and each sample in T in the attribute space. However, when ϵ perturbation is applied to x , the distance changes and, as a result, the K nearest neighbors of x may also change.

Fortunately, the distance in the attribute space is not affected by label-flipping bias in the dataset T , since label-flipping only impacts sample labels, not sample attributes. Thus, in this subsection, we only need to consider the impact of ϵ perturbation of the input x .

The Challenge. Due to ϵ perturbation, a single test input x becomes a potentially-infinite set of inputs $\Delta^\epsilon(x)$. Since our goal is to over-approximate the K nearest neighbors of $\Delta^\epsilon(x)$, the expectation is that, as long as there exists some $x' \in \Delta^\epsilon(x)$ such that a sample input t in T is one of the K nearest neighbors of x' , denoted $t \in T_{x'}^K$, we must include t in the set *overNN*. That is,

$$\bigcup_{x' \in \Delta^\epsilon(x)} T_{x'}^K \subseteq \text{overNN} \subseteq T.$$

However, finding an efficient way of computing *overNN* is a challenging task. As explained before, the naive approach of enumerating $x' \in \Delta^\epsilon(x)$, computing the K nearest neighbors, $T_{x'}^K$, and unionizing all of them would not work. Instead, we need abstraction that is both efficient and accurate enough in practice.

Our solution is that, for each sample t in T , we first analyze the distances between t and all inputs in $\Delta^\epsilon(x)$ symbolically, to compute a lower bound and an upper bound of the distances. Then, we leverage these lower and upper bounds to compute the set *overNN*, which is a superset of samples in T that may become the K nearest neighbors of $\Delta^\epsilon(x)$.

Bounding Distance Between $\Delta^\epsilon(x)$ and t . Assume that $x = (x_1, x_2, \dots, x_D)$ and $t = (t_1, t_2, \dots, t_D)$ are two real-valued vectors in the D -dimensional attribute space. Let $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_D)$, where $\epsilon_i \geq 0$, be the small perturbation. Thus, the perturbed input is $x' = (x'_1, x'_2, \dots, x'_D) = (x_1 + \delta_1, x_2 + \delta_2, \dots, x_D + \delta_D)$, where $\delta_i \in [-\epsilon_i, \epsilon_i]$ for all $i = 1, \dots, D$.

The distance between x and t is a fixed value $d(x, t) = \sqrt{\sum_{i=1}^D (x_i - t_i)^2}$, since both x and the samples t in T are fixed, but the distance between $x' \in \Delta^\epsilon(x)$ and t is a function of $\delta_i \in [-\epsilon_i, \epsilon_i]$, since $\sqrt{\sum_{i=1}^D (x'_i - t_i)^2} = \sqrt{\sum_{i=1}^D (x_i - t_i + \delta_i)^2}$.

For ease of presentation, we define the distance as $d^\epsilon = \sqrt{\sum_{i=1}^D d_i^\epsilon}$, where $d_i^\epsilon = (x_i - t_i + \delta_i)^2$ is the (squared) distance function in the i -th dimension. Then, our goal becomes computing the lower bound, $LB(d^\epsilon)$, and the upper bound, $UB(d^\epsilon)$, in the domain $\delta_i \in [-\epsilon_i, \epsilon_i]$ for all $i = 1, \dots, D$.

Distance Bounds are Compositional. Our first observation is that bounds on the distance d^ϵ as a whole can be computed using bounds in the individual dimensions. To see why this is the case, consider the (square) distance in the i -th dimension, $d_i^\epsilon = (x_i - t_i + \delta_i)^2$, where $\delta_i \in [-\epsilon_i, \epsilon_i]$, and the (square) distance in the j -th dimension, $d_j^\epsilon = (x_j - t_j + \delta_j)^2$, where $\delta_j \in [-\epsilon_j, \epsilon_j]$. By definition, d_i^ϵ is completely independent of d_j^ϵ when $i \neq j$.

Thus, the lower bound of d^ϵ , denoted $LB(d^\epsilon)$, can be calculated by finding the lower bound of each d_i^ϵ in the i -th dimension. Similarly, the upper bound of d^ϵ , denoted $UB(d^\epsilon)$, can also be calculated by finding the upper bound of each d_i^ϵ in the i -th dimension. That is,

$$LB(d^\epsilon) = \sqrt{\sum_{i=1}^D LB(d_i^\epsilon)} \text{ and } UB(d^\epsilon) = \sqrt{\sum_{i=1}^D UB(d_i^\epsilon)}.$$

Four Cases in Each Dimension. Our second observation is that, by utilizing the mathematical nature of the (square) distance function, we can calculate the minimum and maximum values of d_i^ϵ , which can then be used as the lower bound $LB(d_i^\epsilon)$ and upper bound $UB(d_i^\epsilon)$, respectively.

Specifically, in the i -th dimension, the (square) distance function $d_i^\epsilon = ((x_i - t_i) + \delta_i)^2$ may be rewritten to $(\delta_i + A)^2$, where $A = (x_i - t_i)$ is a constant and $\delta_i \in [-\epsilon, +\epsilon]$ is a variable. The function can be plotted in two dimensional space,

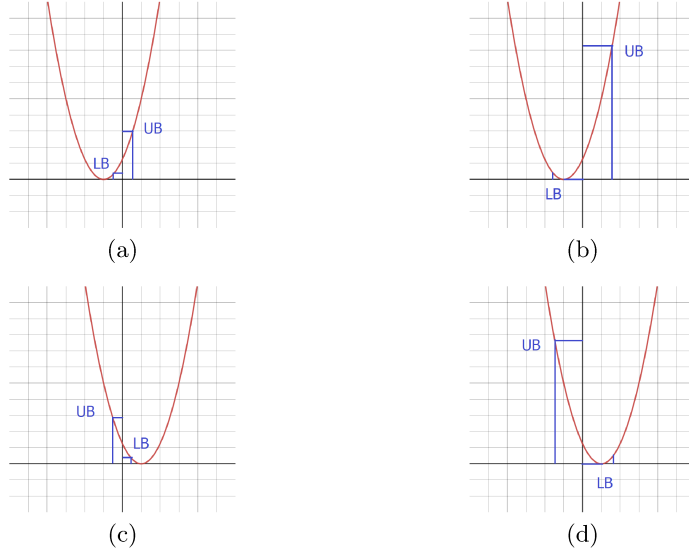


Fig. 3. Four cases for computing the upper and lower bounds of the distance function $d_i^\epsilon(\delta_i) = (\delta_i + A)^2$ for $\delta_i \in [-\epsilon_i, \epsilon_i]$. In these figures, δ_i is the x -axis, and d_i^ϵ is the y -axis, LB denotes $LB(d_i^\epsilon)$, and UB denotes $UB(d_i^\epsilon)$.

using δ_i as x -axis and the output of the function as y -axis; thus, it is a quadratic function $Y = (X + A)^2$.

Fig. 3 shows the plot, which reminds us of where the minimum and maximum values of a quadratic function is. There are two versions of the quadratic function, depending on whether $A > 0$ (corresponding to the two subfigures at the top) or $A < 0$ (corresponding to the two subfigures at the bottom). Each version also has two cases, depending on whether the perturbation interval $[-\epsilon_i, \epsilon_i]$ falls inside the constant interval $[-|A|, |A|]$ (corresponding to the two subfigures on the left) or falls outside (corresponding to the two subfigures on the right). Thus, there are four cases in total.

In each case, the maximal and minimal values of the quadratic function are different, as shown by the LB and UB marks in Fig. 3.

Case (a) This is when $(x_i - t_i) > 0$ and $-\epsilon_i > -(x_i - t_i)$, which is the same as saying $A > 0$ and $-\epsilon_i > -A$. In this case, function $d_i(\epsilon_i) = (\delta_i + A)^2$ is monotonically increasing w.r.t. variable $\delta_i \in [-\epsilon_i, +\epsilon_i]$.

Thus, $LB(d_i^\epsilon) = (-\epsilon_i + (x_i - t_i))^2$ and $UB(d_i^\epsilon) = (+\epsilon_i + (x_i - t_i))^2$.

Case (b) This is when $(x_i - t_i) > 0$ and $-\epsilon_i < -(x_i - t_i)$, which is the same as saying $A > 0$ and $-\epsilon_i < -A$. In this case, the function is not monotonic. The minimal value is 0, obtained when $\delta_i = -A$. The maximal value is obtained when $\delta_i = +\epsilon_i$.

Thus, $LB(d_i^\epsilon) = 0$ and $UB(d_i^\epsilon) = (+\epsilon_i + (x_i - t_i))^2$.

Case (c) This is when $(x_i - t_i) < 0$ and $\epsilon_i < -(x_i - t_i)$, which is the same as saying $A < 0$ and $\epsilon_i < -A$. In this case, the function is monotonically decreasing w.r.t. variable $\delta_i \in [-\epsilon_i, \epsilon_i]$.

Thus, $LB(d_i^\epsilon) = (\epsilon_i + (x_i - t_i))^2$ and $UB(d_i^\epsilon) = (-\epsilon_i + (x_i - t_i))^2$.

Case (d) This is when $(x_i - t_i) < 0$ and $\epsilon_i > -(x_i - t_i)$, which is the same as saying $A < 0$ and $\epsilon_i > -A$. In this case, the function is not monotonic. The minimal value is 0, obtained when $\delta_i = -A$. The maximal value is obtained when $\delta_i = -\epsilon_i$.

Thus, $LB(d_i^\epsilon) = 0$ and $UB(d_i^\epsilon) = (-\epsilon_i + (x_i - t_i))^2$.

Summary By combining the above four cases, we compute the bounds of the entire distance function d^ϵ as follows:

$$\left[\sqrt{\sum_{i=1}^D \max(|x_i - t_i| - \epsilon_i, 0)^2}, \sqrt{\sum_{i=1}^D (|x_i - t_i| + \epsilon_i)^2} \right]$$

Here, the take-away message is that, since x_i , t_i and ϵ_i are all fixed values, the upper and lower bounds can be computed in constant time, despite that there is a potentially-infinite number of inputs in $\Delta^\epsilon(x)$.

Computing *overNN* Using Bounds. With the upper and lower bounds of the distance between $\Delta^\epsilon(x)$ and sample t in the dataset T , denoted $[LB(d^\epsilon(x, t)), UB(d^\epsilon(x, t))]$, we are ready to compute *overNN* such that every $t \in \text{overNN}$ may be among the K nearest neighbors of $\Delta^\epsilon(x)$.

Let UB_{Kmin} denote the K -th minimum value of $UB(d^\epsilon(x, t))$ for all $t \in T$. Then, we define *overNN* as the set of samples in T whose $LB(d^\epsilon(x, t))$ is *not greater than* UB_{Kmin} . In other words,

$$\text{overNN} = \{t \in T \mid LB(d^\epsilon(x, t)) \leq UB_{Kmin}\}.$$

Example Given a dataset $T = \{t^1, t^2, t^3, t^4, t^5\}$, a test input x , perturbation ϵ , and $K = 3$. Assume that the lower and upper bounds of the distance between $\Delta^\epsilon(x)$ and samples in T are $[25.4, 29.4]$, $[30.1, 34.1]$, $[35.3, 39.3]$, $[37.2, 41.2]$, $[85.5, 90.5]$. Since $K = 3$, we find the 3rd minimum upper bound, $UB_{3min} = 39.3$. By comparing UB_{3min} with the lower bounds, we compute $\text{overNN}_3 = \{t^1, t^2, t^3, t^4\}$, since t^5 is the only sample in T whose lower bound is greater than 39.3. All the other four samples may be among the 3 nearest neighbors of $\Delta^\epsilon(x)$.

Due to ϵ perturbation, the set overNN_3 for $K = 3$ is expected to contain 3 or more samples. That is, since different inputs in $\Delta^\epsilon(x)$ may have different samples as their 3-nearest neighbors, to be conservative, we have to take the union of all possible sets of 3-nearest neighbors.

Algorithm 2: Subroutine `abs_same_label( $overNN, K, y$ )`.

```

1 Let  $S$  be a subset of  $overNN$  obtained by removing all  $y$ -labeled elements;
2 Let  $y' = \text{Freq}(S)$ , and  $\#y'$  be the count of  $y'$ -labeled elements in  $S$ ;
3 if  $\#y' < K - |S| - 2 * n$  then
4   | return True;
5 end if
6 return False;
    
```

Soundness Proof Here we prove that any $t' \notin overNN_K$ cannot be among the K nearest neighbors of any $x' \in \Delta^\epsilon(x)$. Since UB_{Kmin} is the K -th minimum $UB(d^\epsilon(x, t))$ for all $t \in T$, there must be samples $t^1, t^2, \dots, t^K$ such that $UB(d^\epsilon(x, t^i)) \leq UB_{Kmin}$ for all $i = 1, 2, \dots, K$. For any $t' \notin overNN$, we have $LB(d^\epsilon(x, t')) > UB_{Kmin}$.

Combining the above conditions, we have $LB(d^\epsilon(x, t')) > UB(d^\epsilon(x, t^i))$ for $i = 1, 2, \dots, K$. It means at least K other samples are closer to x than t' . Thus, t' cannot be among the K -nearest neighbors of x' .

4.2 Checking the Classification Result

Next, we try to certify that, regardless of which of the K elements are selected from $overNN$, the prediction result obtained using them is always the same.

The prediction label is affected by both ϵ perturbation of the input x and label-flipping bias in the dataset T . Since ϵ perturbation affects which points are identified as the K nearest neighbors, and its impact has been accounted for by $overNN$, from now on, we focus only on label-flipping bias in T .

Our method is shown in Algorithm 2, which takes the set $overNN$, the parameter K , and the expected label y as input, and checks if it is possible to find a subset of $overNN$ with size K , whose most frequent label differs from y . If such a “bad” subset cannot be found, we say that KNN prediction always returns the same label.

To try to find such a “bad” subset of $overNN$, we first remove all elements labeled with y from $overNN$, to obtain the set S (Line 1). After that, there are two cases to consider.

1. If the size of S is equal to or greater than K , then any subset of S with size K must have a different label because it will not contain any element labeled with y . Thus, the condition in Line 3 of Algorithm 2 is not satisfied ($\#y'$ is a positive number, and right-hand side is a negative number), and the procedure returns *False*.
2. If the size of S , denoted $|S|$, is smaller than K , the most likely “bad” subset will be $S_K = S \cup \{\text{any } (K - |S|) \text{ } y\text{-labeled elements from } overNN\}$. In this case, we need to check if the most frequent label in S_K is y or not.

In S_K , the most frequent label must be either y (whose count is $K - |S|$) or y' (which is the most frequent label in S , with the count $\#y'$). Moreover, since we can flip up to n labels, we can flip n elements from label y to label y' .

Algorithm 3: Subroutine *abs_KNN_learn*(T, n)

```

1 for each candidate  $k$  value do
2   Let  $\{G_i\}$  = a partition of  $T$  into  $p$  groups of roughly equal size;
3    $errUB_i^k = \{(x, y) \in G_i \mid \text{abs\_may\_err}(T \setminus G_i, n, k, x, y) = \text{true}\}$  for each
       $G_i$ ;
4    $errLB_i^k = \{(x, y) \in G_i \mid \text{abs\_must\_err}(T \setminus G_i, n, k, x, y) = \text{true}\}$  for each
       $G_i$ ;
5    $UB_k = \frac{1}{p} \sum_{i=1}^p |errUB_i^k|/|G_i|$ ;
6    $LB_k = \frac{1}{p} \sum_{i=1}^p |errLB_i^k|/|G_i|$ ;
7 end for
8 Let  $minUB = \min(\{UB_1, \dots, UB_p\})$ ;
9 return  $KSet = \{k \mid LB_k \leq minUB\}$ ;

```

Therefore, to check if our method should return *True*, meaning the prediction result is guaranteed to be the same as label y , we only need to compare $K - |S|$ with $\#y' + 2 * n$. This is checked using the condition in Line 3 of Algorithm 2.

5 Abstracting the KNN Learning Step

In this section, we present our method for abstracting the learning step, which computes the optimal K value based on T and the impact of flipping at most n labels. The output is a super set of possible optimal K values, denoted $KSet$.

Algorithm 3 shows our method, which takes the training set T and parameter n as input, and returns $KSet$ as output. To be sound, we require the $KSet$ to include any candidate k value that may become the optimal K for some clean set $T' \in \text{dBias}_n(T)$.

In Algorithm 3, our method first computes the lower and upper bounds of the classification error for each k value, denoted LB_k and UB_k , as shown in Lines 5-6. Next, it computes $minUB$, which is the minimal upper bound for all candidate k values (Line 8). Finally, by comparing $minUB$ with LB_k for each candidate k value, our method decides whether this candidate k value should be put into $KSet$ (Line 9).

We will explain the steps needed to compute LB_k and UB_k in the remainder of this section. For now, assuming that they are available, we explain how they are used to compute $KSet$.

Example Given the candidate k values, k_1, k_2, k_3, k_4 , and their error bounds $[0.1, 0.2]$, $[0.1, 0.3]$, $[0.3, 0.4]$, $[0.3, 0.5]$. The smallest upper bound is $minUB = 0.2$. By comparing $minUB$ with the lower bounds, we compute $KSet = \{k_1, k_2\}$, since only LB_{k_1} and LB_{k_2} are lower than or equal to $minUB$.

Soundness Proof Here we prove that any $k' \notin KSet$ cannot result in the smallest classification error. Assume that k_s is the candidate k value that has the minimal upper bound ($minUB$), and err_{k_s} is the actual classification error. By definition,

Algorithm 4: Subroutine `abs_may_err`(T, n, K, x, y).

- 1 Let y' be, among the non- y labels, the label with the highest count in T_x^K ;
 - 2 Let $\#y$ be the number of elements in T_x^K with the y label;
 - 3 Let n' be $\min(n, \#y \in T_x^K)$;
 - 4 Changing n' elements in T_x^K from y label to y' label;
 - 5 **return** $\text{Freq}(T_x^K) \neq y$;
-

we have $\text{err}_{k_s} \leq \text{minUB}$. Meanwhile, for any $k' \notin K\text{Set}$, we have $LB_{k'} > \text{minUB}$. Combining the two cases, we have $\text{err}_{k'} > \text{minUB} \geq \text{err}_{k_s}$. Here, $\text{err}_{k'} > \text{err}_{k_s}$ means that k' cannot result in the smallest classification error.

5.1 Overapproximating the Classification Error

To compute the upper bound errUB_i^k defined in Line 3 of Algorithm 3, we use the subroutine `abs_may_err` to check if $(x, y) \in G_i$ may be misclassified when using $T \setminus G_i$ as the training set.

Algorithm 4 shows the implementation of the subroutine, which checks, for a sample (x, y) , whether it is possible to obtain a set S by flipping at most n labels in T_x^K such that the most frequent label in S is not y . If it is possible to obtain such a set S , we conclude that the prediction label for x may be an error.

The condition $\text{Freq}(T_x^K) \neq y$, computed on T_x^K after the y label of n' elements is changed to y' label, is a sufficient condition under which the prediction label for x may be an error. The rationale is as follows.

In order to make the most frequent label in the set T_x^K different from y , we need to focus on the label most likely to become the new most frequent label. It is the label $y' (\neq y)$ with the highest count in the current T_x^K .

Therefore, Algorithm 4 checks whether y' can become the most frequent label by changing at most n elements in T_x^K from y label to y' label (Lines 3-5).

5.2 Underapproximating the Classification Error

To compute the lower bound errLB_i^k defined in Line 4 of Algorithm 3, we use the subroutine `abs_must_err` to check if $(x, y) \in G_i$ must be misclassified when using $T \setminus G_i$ as the training set.

Algorithm 5 shows the implementation of the subroutine, which checks, for a sample (x, y) , whether it is impossible to obtain a set S by flipping at most n labels in T_x^K such that the most frequent label in S is y . In other words, is it impossible to avoid the classification error? If it is impossible to avoid the classification error, we conclude that the prediction label must be an error, and thus the procedure returns *True*.

In this sense, all samples in errLB_i^k (computed in Line 4 of Algorithm 3) are guaranteed to be misclassified.

Algorithm 5: Subroutine `abs_must_err`(T, n, K, x, y).

```

1 if  $\exists S$  obtained from  $T_x^K$  by flipping up to  $n$  labels such that  $\text{Freq}(S) = y$  then
2   | return False;
3 end if
4 return True;

```

The challenge in Algorithm 5 is to check if such a set S can be constructed from T_x^K . The intuition is that, to make y the most frequent label, we should flip the labels of non- y elements to label y . Let us consider two examples first.

Example 1 Given the label counts of T_x^K , denoted $\{l_1 * 4, l_4 * 4, l_3 * 2\}$, meaning that 4 elements are labeled l_1 , 4 elements are labeled l_4 , and 2 elements are labeled l_3 . Assume that $n = 2$ and $y = l_3$. Since we can flip at most 2 elements, we choose to flip one $l_1 \rightarrow l_3$ and one $l_4 \rightarrow l_3$, to get a set $S = \{l_1 * 3, l_4 * 3, l_3 * 4\}$.

Example 2 Given the label counts of T_x^K , denoted $\{l_1 * 5, l_4 * 3, l_3 * 2\}$, $n = 2$, and $y = l_3$. We can flip two $l_1 \rightarrow l_3$ to get a set $S = \{l_1 * 3, l_4 * 3, l_3 * 4\}$.

The LP Problem The question is how to decide whether the set S (defined in Line 1 of Algorithm 5) exists. We can formulate it as a linear programming (LP) problem. The LP problem has two constraints. The first one is defined as follows: Let y be the expected label, $l_i \neq y$ be another label, where $i = 1, \dots, q$ and q is the total number of class labels (e.g., in the above two examples, the number $q = 3$). Let $\#y$ be the number of elements in T_x^K that have the y label. Similarly, let $\#l_i$ be the number of elements with l_i label. Assume that a set S as defined in Algorithm 5 exists, then all of the labels $l_i \neq y$ must satisfy

$$\#l_i - \#flip_i < \#y + \sum_{i=1}^q \#flip_i, \quad (1)$$

where $\#flip_i$ is a variable representing the number of l_i -to- y flips. Thus, in the above formula, the left-hand side is the count of l_i after flipping, the right-hand side is the count of y after flipping. Since y is the most frequent label in S , y should have a higher count than any other label.

The second constraint is

$$\sum_{i=1}^q \#flip_i \leq n, \quad (2)$$

which says that the total number of label flips is bounded by the parameter n .

Since the number of class labels (q) is often small (from 2 to 10), this LP problem can be solved quickly. However, the LP problem must be solved $|T|$ times, where $|T|$ may be as large as 50,000. To avoid invoking the LP solver

unnecessarily, we propose two easy-to-check conditions. They are *necessary* condition in that, if either of them is violated, the set S does not exist. Thus, we invoke the LP solver only if both conditions are satisfied.

Necessary Conditions The first condition is derived from Formula (1a), by adding up the two sides of the inequality constraint for all labels $l_i \neq y$. The resulting condition is

$$\left(\sum_{l_i \neq y} \#l_i - \sum_{i=1}^q \#flip_i \right) < \left((q-1)\#y + (q-1) \sum_{i=1}^q \#flip_i \right).$$

The second condition requires that, in S , label y has a higher count (after flipping) than any other label, including the label $l_p \neq y$ with the highest count in the current T_x^K . The resulting condition is

$$(\#l_p - \#y)/2 < n,$$

since only when this condition is satisfied, it is possible to allow y to have a higher count than l_p , by flipping at most n of the label l_p to y .

These are necessary conditions (but may not be sufficient conditions) because, whenever the first condition does not hold, Equation (1) does not hold either. Similarly, whenever the second condition does not hold, Equation (1) does not hold either. In this sense, these two conditions are *easy-to-check* over-approximations of Equation (1).

6 Experiments

We have implemented our method as a software tool written in Python using the **scikit-learn** machine learning library. We evaluated our tool on six datasets that are widely used in the fairness research literature.

Datasets Table 1 shows the statistics of each dataset, including the name, a short description, the size ($|T|$), the number of attributes, the protected attributes, and the parameters ϵ and n . The value of ϵ is set to 1% of the attribute range. The bias parameter n is set to 1 for small datasets, 10 for medium datasets, and 50 for large datasets. The protected attributes include *Gender* for all six datasets, and *Race* for two datasets, *Compas* and *Adult*, which are consistent with known biases in these datasets.

In preparation for the experimental evaluation, we have employed state-of-the-art techniques in the machine learning literature to preprocess and balance the datasets for KNN, including encoding, standard scaling, k-bins-discretizer, downsampling and upweighting.

Table 1. Statistics of all of the datasets used during our experimental evaluation.

Dataset	Description	Size $ T $	# Attr.	Protected Attr.	Parameters ϵ and n
Salary	salary level [42]	52	4	Gender	$\epsilon = 1\%$ attribute range, $n = 1$
Student	academic performance [9]	649	30	Gender	$\epsilon = 1\%$ attribute range, $n = 1$
German	credit risk [13]	1,000	20	Gender	$\epsilon = 1\%$ attribute range, $n = 10$
Compas	recidivism risk [11]	10,500	16	Race+Gender	$\epsilon = 1\%$ attribute range, $n = 10$
Default	loan default risk [47]	30,000	36	Gender	$\epsilon = 1\%$ attribute range, $n = 50$
Adult	earning power [13]	48,842	14	Race+Gender	$\epsilon = 1\%$ attribute range, $n = 50$

Table 2. Results for certifying *label-flipping* and *individual fairness* (gender) on small datasets, for which ground truth can still be obtained by naive enumeration, and compared with our method.

Name	Certifying label-flipping fairness						Certifying label-flipping + individual fairness					
	Ground truth	Time	Our method	Time	Accuracy	Speedup	Ground truth	Time	Our method	Time	Accuracy	Speedup
Salary	50.0%	1.7s	33.3%	0.2s	66.7%	8.5X	33.3%	1.5s	33.3%	0.2s	100%	7.5X
Student	70.8%	23.0s	60.0%	0.2s	84.7%	115X	58.5%	25.2s	44.6%	0.2s	76.2%	116X

Methods For comparison purposes, we implemented six variants of our method, by enabling or disabling the ability to certify label-flipping fairness, the ability to certify individual fairness, and the ability to certify ϵ -fairness.

Except for ϵ -fairness, we also implemented the naive approach of enumerating all $T' \in \text{dBias}_n(T)$. Since the naive approach does not rely on approximation, its result can be regarded as the ground truth (i.e., whether the classification output for an input x is truly fair). Our goal is to obtain the ground truth on small datasets, and use it to evaluate the accuracy of our abstract interpretation based method. However, as explained before, enumeration does not work for ϵ -fairness, since the number of inputs in $\Delta^\epsilon(x)$ is infinite.

Our experiments were conducted on a computer with 2 GHz Quad-Core Intel Core i5 CPU and 16 GB of memory. The experiments were designed to answer two questions. First, is our method efficient and accurate enough in handling popular datasets in the fairness literature? Second, does our method help us gain insights? For example, it would be interesting to know whether decision made on an individuals from a protected minority group is more (or less) likely to be certified as fair.

Results on Efficiency and Accuracy We first evaluate the efficiency and accuracy of our method. For the two small datasets, *Salary* and *Student*, we are able to obtain the ground truth using the naive enumeration approach, and then compare it with the result of our abstract interpretation based method. We want to know how much our results deviate from the ground truth.

Table 2 shows the results obtained by treating *Gender* as the protected attribute. Column 1 shows the name of the dataset. Columns 2-7 compare the naive approach (ground truth) and our method in certifying label-flipping fairness. Columns 8-13 compare the naive approach (ground truth) and our method in certifying label-flipping plus individual fairness.

Table 3. Results for certifying *label-flipping*, *individual*, and ϵ -*fairness* by our method.

Name	Label-flipping fairness	Time	+ Individual fairness	Time	+ ϵ -fairness	Time
Salary (gender)	33.3%	0.2s	33.3%	0.2s	33.3%	0.2s
Student (gender)	60.0%	0.2s	44.6%	0.2s	32.3%	0.2s
German (gender)	48.0%	0.2s	44.0%	0.3s	43.0%	0.2s
Compas (race)	95.0%	0.3s	62.4%	1.4s	56.4%	1.1s
Compas (gender)	95.0%	0.3s	65.3%	1.3s	59.4%	1.0s
Default (gender)	83.2%	2.3s	73.3%	4.4s	64.4%	3.5s
Adult (race)	76.2%	2.2s	65.3%	4.5s	53.5%	5.3s
Adult (gender)	76.2%	2.2s	52.5%	3.5s	43.6%	3.3s

Based on the results in Table 2, we conclude that the accuracy of our method is high (81.9% on average) despite its aggressive use of abstraction to reduce the computational cost. Our method is also 7.5X to 126X faster than the naive approach. Furthermore, the larger the dataset, the higher the speedup.

For medium and large datasets, it is infeasible for the naive enumeration approach to compute and show the ground truth in Table 2. However, the fairness scores of our method shown in Table 3 provide “lower bounds” for the ground truth since our method is sound for certification. For example, when our method reports 95% for *Compas (race)* in Table 3, it means the ground truth must be $\geq 95\%$ (and thus the gap must be $\leq 5\%$). However, there does not seem to be obvious relationship between the gap and the dataset size – the gap may be due to some unique characteristics of each dataset.

Results on the Certification Rates We now present the success rates of our certification method for the three variants of fairness. Table 3 shows the results for label-flipping fairness in Columns 2-3, label-flipping plus individual fairness (denoted *+ Individual fairness*) in Columns 4-5, and label-flipping plus ϵ -fairness (denoted *+ ϵ -fairness*) in Columns 6-7. For each variant of fairness, we show the percentage of test inputs that are certified to be fair, together with the average certification time (per test input). In all six datasets, *Gender* was treated as the protected attribute. In addition, *Race* was treated as the protected attribute for *Compas* and *Adult*.

From the results in Table 3, we see that as more stringent fairness standard is used, the certified percentage either stays the same (as in *Salary*) or decreases (as in *Student*). This is consistent with what we expect, since the classification output is required to stay the same for an increasingly larger number of scenarios. For *Compas (race)*, in particular, adding ϵ -fairness on top of label-flipping fairness causes the certified percentage to drop from 62.4% to 56.4%.

Nevertheless, our method still maintains a high certification percentage. Recall that, for *Salary*, the 33.3% certification rate (for *+ Individual fairness*) is actually 100% accurate according to comparison with the ground truth in Table 2, while the 44.6% certification rate (for *+ Individual fairness*) is actually 76.2% accurate. Furthermore, the efficiency of our method is high: for *Adult*, which has 50,000 samples in the training set, the average certification time of our method remains within a few seconds.

Table 4. Results for certifying *label-flipping* + ϵ -*fairness* with both *Race* and *Gender* as protected attributes.

	White	Other	Wt. Avg
(a) Compas			
Male	61.9%	52.2%	52.8%
Female	100%	60.0%	63.7%
Wt. Avg	63.7%	53.7%	54.4%

	White	Other	Wt. Avg
(b) Adult			
Male	35.3%	33.3%	35.1%
Female	33.3%	66.7%	37.0%
Wt. Avg	34.7%	44.4%	35.6%

Results on Demographic Groups Table 4 shows the certified percentage of each demographic group, when both *label-flipping* and ϵ -*fairness* are considered, and both *Race* and *Gender* are treated as protected attributes. The four demographic groups are (1) *White Male*, (2) *White Female*, (3) *Other Male*, and (4) *Other Female*. For each group, we show the certified percentage obtained by our method. In addition, we show the weighted averages for *White* and *Other*, as well as the weighted averages for *Male* and *Female*.

For *Compas*, *White Female* has the highest certified percentage (100%) while *Other Female* has the lowest certified percentage (52.2%); here, the classification output represents the recidivism risk.

For *Adult*, *Other Female* has the highest certified percentage (66.7%) while the other three groups have certified percentages in the range of 33.3%-35.3%.

The differences may be attributed to two sources, one of which is technical and the other is social. The social reason is related to historical bias, which is well documented for these datasets. If the actual percentages (ground truth) is different, the percentages reported by our method will also be different. The technical reason is related to the nature of the KNN algorithm itself, which we explain as follows.

In these datasets, some demographic groups have significantly more samples than others. In KNN, the lowest occurring group may have a limited number of close neighbors. Thus, for each test input x from this group, its K nearest neighbors tend to have a larger radius in the input vector space. As a result, the impact of ϵ perturbation on x will be smaller, resulting in fewer changes to its K nearest neighbors. That may be one of the reasons why, in Table 4, the lowest occurring groups, *White Female* in *Compas* and *Other Female* in *Adult*, have significantly higher certified percentage than other groups.

Results in Table 4 show that, even if a machine learning technique discriminates against certain demographic groups, for an individual, the prediction result produced by the machine learning technique may still be fair. This is closely related to differences (and sometimes conflicts) between *group fairness* and *individual fairness*: while group fairness focuses on statistical parity, individual fairness focuses on similar outcomes for similar individuals. Both are useful notions and in many cases they are complementary.

Caveat Our work should not be construed as an endorsement nor criticism of the use of machine learning techniques in socially sensitive applications. Instead, it should be viewed as an effort on developing new methods and tools to help improve our understanding of these techniques.

7 Related Work

For fairness certification, as explained earlier in this paper, our method is the first method for certifying KNN in the presence of historical (dataset) bias. While there are other KNN certification and falsification techniques, including Jia et al. [20] and Li et al. [24, 25], they focus solely on robustness against data poisoning attacks as opposed to individual and ϵ -fairness against historical bias. Meyer et al. [26, 27] and Drews et al. [12] propose certification techniques that handle dataset bias, but target different machine learning techniques (decision tree or linear regression); furthermore, they do not handle ϵ -fairness.

Throughout this paper, we have assumed that the KNN learning (parameter-tuning) step is not tampered with or subjected to fairness violation. However, since the only impact of tampering with the KNN learning step will be changing the optimal value of the parameter K , the biased KNN learning step can be modeled using a properly over-approximated $KSet$. With this new $KSet$, our method for certifying fairness of the prediction result (as presented in Section 4) will work AS IS.

Our method aims to certify fairness with certainty. In contrast, there are statistical techniques that can be used to prove that a system is fair or robust with a high probability. Such techniques have been applied to various machine learning models, for example, in *VeriFair* [6] and *FairSquare* [2]. However, they are typically applied to the prediction step while ignoring the learning step, although the learning step may be affected by dataset bias.

There are also techniques for mitigating bias in machine learning systems. Some focus on improving the learning algorithms using random smoothing [33], better embedding [7] or fair representation [34], while others rely on formal methods such as iterative constraint solving [38]. There are also techniques for repairing models to improve fairness [3]. Except for Ruoss et al. [34], most of them focus on group fairness such as demographic parity and equal opportunity; they are significantly different from our focus on certifying individual and ϵ -fairness of the classification results in the presence of dataset bias.

At a high level, our method that leverages a sound over-approximate analysis to certify fairness can be viewed as an instance of the abstract interpretation paradigm [10]. Abstract interpretation based techniques have been successfully used in many other settings, including verification of deep neural networks [17, 30], concurrent software [21, 22, 37], and cryptographic software [43, 44].

Since fairness is a type of non-functional property, the verification/certification techniques are often significantly different from techniques used to verify/certify functional correctness. Instead, they are more closely related to techniques for verifying/certifying robustness [8], noninterference [5], and side-channel security [19, 39, 40, 48], where a program is executed multiple times, each time for a different input drawn from a large (and sometimes infinite) set, to see if they all agree on the output. At a high level, this is closely related to differential verification [28, 31, 32], synthesis of relational invariants [41] and verification of hyper-properties [15, 35].

8 Conclusions

We have presented a method for certifying the individual and ϵ -fairness of the classification output of the KNN algorithm, under the assumption that the training dataset may have historical bias. Our method relies on abstract interpretation to soundly approximate the arithmetic computations in the learning and prediction steps. Our experimental evaluation shows that the method is efficient in handling popular datasets from the fairness research literature and accurate enough in obtaining certifications for a large amount of test data. While this paper focuses on KNN only, as a future work, we plan to extend our method to other machine learning models.

References

1. Adeniyi, D.A., Wei, Z., Yongquan, Y.: Automated web usage data mining and recommendation system using k-nearest neighbor (knn) classification method. *Applied Computing and Informatics* **12**(1), 90–108 (2016)
2. Albarghouthi, A., D’Antoni, L., Drews, S., Nori, A.V.: Fairsquare: probabilistic verification of program fairness. *Proceedings of the ACM on Programming Languages* **1**(OOPSLA), 1–30 (2017)
3. Albarghouthi, A., D’Antoni, L., Drews, S.: Repairing decision-making programs under uncertainty. In: *International Conference on Computer Aided Verification*. pp. 181–200. Springer (2017)
4. Andersson, M., Tran, L.: Predicting movie ratings using KNN (2020)
5. Barthe, G., D’Argenio, P.R., Rezk, T.: Secure information flow by self-composition. In: *IEEE Computer Security Foundations Workshop*. pp. 100–114 (2004)
6. Bastani, O., Zhang, X., Solar-Lezama, A.: Probabilistic verification of fairness properties via concentration. *Proceedings of the ACM on Programming Languages* **1**(OOPSLA), 1–27 (2019)
7. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Annual Conference on Neural Information Processing Systems* **29** (2016)
8. Chaudhuri, S., Gulwani, S., Lubliner, R.: Continuity and robustness of programs. *Commun. ACM* **55**(8), 107–115 (2012)
9. Cortez, P., Silva, A.M.G.: Using data mining to predict secondary school student performance. *EUROSIS-ETI* (2008)
10. Cousot, P., Cousot, R.: Abstract interpretation: A unified lattice model for static analysis of programs by construction or approximation of fixpoints. In: *ACM Symposium on Principles of Programming Languages*. pp. 238–252 (1977)
11. Dieterich, W., Mendoza, C., Brennan, T.: COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Northpointe Inc (2016)
12. Drews, S., Albarghouthi, A., D’Antoni, L.: Proving data-poisoning robustness in decision trees. In: *ACM SIGPLAN International Conference on Programming Language Design and Implementation*. pp. 1083–1097 (2020)
13. Dua, D., Graff, C.: UCI machine learning repository (2017), <http://archive.ics.uci.edu/ml>
14. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.S.: Fairness through awareness. In: *Innovations in Theoretical Computer Science*. pp. 214–226 (2012)

15. Finkbeiner, B., Haas, L., Torfah, H.: Canonical representations of k-safety hyperproperties. In: IEEE Computer Security Foundations Symposium. pp. 17–31 (2019)
16. Firdausi, I., Erwin, A., Nugroho, A.S., et al.: Analysis of machine learning techniques used in behavior-based malware detection. In: 2010 second international conference on advances in computing, control, and telecommunication technologies. pp. 201–203. IEEE (2010)
17. Gehr, T., Mirman, M., Drachsler-Cohen, D., Tsankov, P., Chaudhuri, S., Vechev, M.T.: AI2: safety and robustness certification of neural networks with abstract interpretation. In: IEEE Symposium on Security and Privacy. pp. 3–18 (2018)
18. Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K.: KNN model-based approach in classification. In: OTM Confederated International Conferences” On the Move to Meaningful Internet Systems”. pp. 986–996. Springer (2003)
19. Guo, S., Wu, M., Wang, C.: Adversarial symbolic execution for detecting concurrency-related cache timing leaks. In: ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. pp. 377–388 (2018)
20. Jia, J., Liu, Y., Cao, X., Gong, N.Z.: Certified robustness of nearest neighbors against data poisoning and backdoor attacks. In: The AAAI Conference on Artificial Intelligence (2022)
21. Kusano, M., Wang, C.: Flow-sensitive composition of thread-modular abstract interpretation. In: ACM SIGSOFT International Symposium on Foundations of Software Engineering. pp. 799–809 (2016)
22. Kusano, M., Wang, C.: Thread-modular static analysis for relaxed memory models. In: ACM Joint Meeting on European Software Engineering Conference and Symposium on Foundations of Software Engineering. pp. 337–348 (2017)
23. Li, Y., Fang, B., Guo, L., Chen, Y.: Network anomaly detection based on TCM-KNN algorithm. In: ACM symposium on Information, Computer and Communications Security. pp. 13–19 (2007)
24. Li, Y., Wang, J., Wang, C.: Proving robustness of KNN against adversarial data poisoning. In: International Conference on Formal Methods in Computer-Aided Design. pp. 7–16 (2022)
25. Li, Y., Wang, J., Wang, C.: Systematic testing of the data-poisoning robustness of KNN. In: ACM SIGSOFT International Symposium on Software Testing and Analysis (2023)
26. Meyer, A.P., Albarghouthi, A., D’Antoni, L.: Certifying robustness to programmable data bias in decision trees. In: Annual Conference on Neural Information Processing Systems. pp. 26276–26288 (2021)
27. Meyer, A.P., Albarghouthi, A., D’Antoni, L.: Certifying data-bias robustness in linear regression. CoRR **abs/2206.03575** (2022)
28. Mohammadinejad, S., Paulsen, B., Deshmukh, J.V., Wang, C.: DiffRNN: Differential verification of recurrent neural networks. In: International Conference on Formal Modeling and Analysis of Timed Systems. pp. 117–134 (2021)
29. Narudin, F.A., Feizollah, A., Anuar, N.B., Gani, A.: Evaluation of machine learning classifiers for mobile malware detection. *Soft Computing* **20**(1), 343–357 (2016)
30. Paulsen, B., Wang, C.: Example guided synthesis of linear approximations for neural network verification. In: International Conference on Computer Aided Verification. pp. 149–170 (2022)
31. Paulsen, B., Wang, J., Wang, C.: ReluDiff: differential verification of deep neural networks. In: International Conference on Software Engineering. pp. 714–726 (2020)

32. Paulsen, B., Wang, J., Wang, J., Wang, C.: NEURODIFF: scalable differential verification of neural networks using fine-grained approximation. In: International Conference on Automated Software Engineering. pp. 784–796 (2020)
33. Rosenfeld, E., Winston, E., Ravikumar, P., Kolter, J.Z.: Certified robustness to label-flipping attacks via randomized smoothing. In: International Conference on Machine Learning. vol. 119, pp. 8230–8241 (2020)
34. Ruoss, A., Balunovic, M., Fischer, M., Vechev, M.T.: Learning certified individually fair representations. In: Annual Conference on Neural Information Processing Systems (2020)
35. Sousa, M., Dillig, I.: Cartesian hoare logic for verifying k-safety properties. In: ACM SIGPLAN Conference on Programming Language Design and Implementation. pp. 57–69 (2016)
36. Su, M.Y.: Real-time anomaly detection systems for denial-of-service attacks by weighted k-nearest-neighbor classifiers. *Expert Systems with Applications* **38**(4), 3492–3498 (2011)
37. Sung, C., Kusano, M., Wang, C.: Modular verification of interrupt-driven software. In: International Conference on Automated Software Engineering. pp. 206–216 (2017)
38. Wang, J., Li, Y., Wang, C.: Synthesizing fair decision trees via iterative constraint solving. In: International Conference on Computer Aided Verification. pp. 364–385. Springer (2022)
39. Wang, J., Sung, C., Raghothaman, M., Wang, C.: Data-driven synthesis of provably sound side channel analyses. In: International Conference on Software Engineering. pp. 810–822 (2021)
40. Wang, J., Sung, C., Wang, C.: Mitigating power side channels during compilation. In: ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. pp. 590–601 (2019)
41. Wang, J., Wang, C.: Learning to synthesize relational invariants. In: International Conference on Automated Software Engineering. pp. 65:1–65:12 (2022)
42. Weisberg, S.: *Applied Linear Regression*, p. 194. John Wiley & Sons (1985)
43. Wu, M., Guo, S., Schaumont, P., Wang, C.: Eliminating timing side-channel leaks using program repair. In: ACM SIGSOFT International Symposium on Software Testing and Analysis. pp. 15–26 (2018)
44. Wu, M., Wang, C.: Abstract interpretation under speculative execution. In: ACM SIGPLAN Conference on Programming Language Design and Implementation. pp. 802–815 (2019)
45. Wu, W., Zhang, W., Yang, Y., Wang, Q.: Drex: Developer recommendation with k-nearest-neighbor search and expertise ranking. In: Asia-Pacific Software Engineering Conference. pp. 389–396 (2011)
46. Xie, M., Hu, J., Han, S., Chen, H.H.: Scalable hypergrid k-nn-based online anomaly detection in wireless sensor networks. *IEEE Transactions on Parallel and Distributed Systems* **24**(8), 1661–1670 (2012)
47. Yeh, I.C., Lien, C.h.: The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications* **36**(2), 2473–2480 (2009)
48. Zhang, J., Gao, P., Song, F., Wang, C.: SCInfer: Refinement-based verification of software countermeasures against side-channel attacks. In: International Conference on Computer Aided Verification. pp. 157–177 (2018)