*algorithms*

MDPI

*Article*

# A Fairness-Aware and Interpretable Model for Recidivism Prediction

**Stamatis Chatzistamatis [1,*], George E. Tsekouras [1], Anastasios Rigos [1], Alvaro Garcia-Recuero [2], Eleni Valari [3], Andreas Siafakas [3] and Konstantinos Kotis [1,*]**

[1] Intelligent Systems Lab, Department of Cultural Technology and Communication, University of the Aegean, 81100 Mytilene, Greece; gtsek@ct.aegean.gr (G.E.T.); a.rigos@aegean.gr (A.R.)
[2] IPS Innovative Prison Systems & ICJT Innovative Criminal Justice Technologies, PARKURBIS Science and Technology Park, 6200-865 Covilhã, Portugal; alvaro.garcia@prisonsystems.eu
[3] IOTAM Ltd.—Internet of Things Applications and Multi Layer Development, 106 Georgiou Griva Digeni Str., 3101 Limassol, Cyprus; e.valari@itml.gr (E.V.); a.siafakas@itml.gr (A.S.)
* Correspondence: stami@aegean.gr (S.C.); kotis@aegean.gr (K.K.)

**Abstract**

Recidivism prediction is increasingly embedded in criminal justice decision-making, yet most deployed systems remain opaque and have been shown to exhibit discriminatory behavior against certain demographic groups. This paper presents a fairness-aware interpretable framework for recidivism prediction applied to three real-world datasets from Bulgaria, Greece, and Portugal. The classification core relies on a 1-Dimensional Convolutional Neural Network (1D-CNN), trained by a custom objective function that embeds the Equalized Odds fairness criterion as an L1-regularized penalty reflecting on gender-based disparities in false positive and false negative rates. Model-level interpretability is provided through Kernel SHAP, which decomposes individual predictions into additive feature attributions grounded in cooperative game theory. Experiments across prediction tasks, each evaluated over randomized runs, demonstrate that the baseline model exhibits statistically significant bias against the female group in all datasets. The fairness-constrained model substantially reduces these disparities across all tasks at a moderate and expected cost to classification accuracy. Kernel SHAP analysis reveals the relative contribution of static and dynamic offenders' attributes to individual risk scores, supporting auditability and contestability. The proposed framework advances the integration of predictive performance, algorithmic fairness, and structural interpretability in criminal justice analytics.

**Keywords:** recidivism prediction; fairness-aware machine learning; Equalized Odds; 1D convolutional neural network; Kernel SHAP; gender bias mitigation; interpretable machine learning; criminal justice analytics; bias mitigation; trustworthy AI

Check for updates

Academic Editor: Paolo Spagnolo

Received: 20 May 2026
Revised: 16 June 2026
Accepted: 20 June 2026
Published: 25 June 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

## 1. Introduction

Recidivism prediction has become one of the most visible applications of data-driven decision support in criminal justice [1–4]. Risk assessment instruments are now used, or actively considered, in pretrial release, sentencing, parole, prison classification, and rehabilitation planning because they promise more consistent, evidence-informed, and scalable decisions than unaided human judgment [1,5,6]. At the same time, the use of predictive systems in criminal justice raises fundamental concerns about fairness, transparency, accountability, and public legitimacy. The relevant literature, therefore, spans doctrinal critiques of risk construction and sentencing use, empirical studies of predictive validity and

bias, technical work on interpretable and fairness-aware modeling, and broader reviews of AI governance in criminal justice [7–11].

The first challenge is that recidivism is not a uniquely defined target. Across the literature, recidivism has been operationalized as rearrest, reconviction, reimprisonment, reincarceration, or a new charge within a specified follow-up window, and these choices materially affect event rates, prediction difficulty, and the meaning of model error [12,13]. Some studies emphasize general post-release reoffending, whereas others focus on violent recidivism, offense-specific recidivism, or return to a particular correctional setting. Recent work further shows that the importance of predictors may vary across one-, two-, three-, and five-year horizons, across adult and juvenile populations, and across general, violent, or offense-specific outcomes [8,14–17]. Consequently, recidivism prediction is not a single fixed learning problem; it is a family of context-dependent prediction tasks shaped by legal definitions, institutional objectives, and operational time frames. This point is especially important for high-stakes applications, because a model optimized for one definition of recidivism may be inappropriate for another, even when the same data source is used.

The field has also evolved substantially in methodological terms. Early actuarial approaches and conventional statistical models used demographic and criminal-history variables to generate simple risk scores, often through additive or linear structures that were readily interpretable but limited in expressive power. Contemporary work has expanded this design space considerably. In that direction, researchers have proposed Bayesian regression and survival models, random forests, support vector machines, gradient boosting systems, decision-tree optimization procedures, neural networks, additive neural models, cluster-aware deep learning pipelines, and fairness-aware multi-objective optimization frameworks [2,18–27]. This diversification reflects two realities. First, recidivism datasets are heterogeneous, frequently imbalanced, and shaped by local correctional practices. Second, there is no universally dominant model family: performance depends on the target definition, data quality, class balance, and institutional context. As a result, recidivism prediction research increasingly focuses not only on whether machine learning can improve discrimination but also on which model is appropriate for a specific criminal justice task.

The modern fairness debate in this area was largely catalyzed by controversies surrounding COMPAS and related tools [10,12,13]. The literature has shown that such systems may produce unequal false-positive and false-negative patterns across demographic groups, especially when evaluated on race and gender subpopulations. Yet, the same literature makes clear that fairness in recidivism prediction is inherently multi-dimensional, as calibration, demographic parity, equal opportunity, predictive equality, predictive parity, and balanced error rates cannot be simultaneously satisfied, especially when outcome-based rates differ across groups [17,28–33]. As a result, there is no single mathematically complete notion of fairness for this problem. Fairness-aware recidivism modeling must instead confront explicit trade-offs among partially incompatible criteria, each reflecting a different normative view about justice, acceptable risk allocation, and the proper role of algorithms in criminal justice.

This recognition moved the field beyond binary debates about whether a model is “fair” or “unfair”. More recent work examines fairness interventions across the full machine learning pipeline, including pre-processing, in-processing, and post-processing stages. Reweighing, adversarial learning, disparate impact removal, reject-option classification, equalized-odds optimization, and related methods have also been applied to recidivism prediction. Integrative studies suggest that isolated debiasing interventions are often insufficient, whereas multi-phase approaches can sometimes significantly improve fairness [33–37]. However, these studies also show that fairness interventions highly depend on the datasets and metrics used. A method that improves one fairness criterion may

worsen another or reduce performance on a different dataset. This means that fairness-aware recidivism prediction is best understood as a constrained or multi-objective-based problem rather than a simple post hoc correction procedure. The implication for model development is direct since fairness should be embedded in the learning objective, model selection strategy, and not treated as an afterthought.

Interpretability forms the second major pillar of current research. In high-stakes structured data settings, several influential studies argue that post hoc explanations of opaque models constitute an insufficient substitute for models that are understandable by design. This concern is reinforced by empirical and conceptual work showing that black-box explanations can be unstable, persuasive without being faithful, or difficult to contest in legally meaningful ways [38–46]. More importantly, interpretability literature does not simply equate interpretability with simplicity. Rather, it emphasizes multiple dimensions, including simulatability, transparency of feature effects, faithful local reasoning, stable global structure, and practical usability for human decision makers. This has led to renewed interest in sparse scoring systems, interpretable classification rules, generalized additive models, explainability-constrained architectures, and neural additive models that selectively incorporate interactions while preserving human-readable component effects [13,40–43,47,48].

A recurring result across the existing literature is that interpretability need not always require a substantial sacrifice in predictive quality. Several comparative studies show that interpretable models can perform as well as black-box alternatives and proprietary tools such as COMPAS or the Arnold PSA when trained on structured recidivism data [5,30,31,42,45]. Other works propose interpretable output constraints, Shapley-based importance analyses across near-optimal models, and explanation architectures that combine global structure with exact local decomposition and counterfactual reasoning [21,39,41]. These contributions have collectively weakened the assumption that accuracy and interpretability lie on opposite ends of a fixed trade-off curve. Instead, they suggest that much depends on whether the prediction task is based on structured tabular data, whether the model class is chosen deliberately, and whether explanation is treated as an intrinsic modeling requirement rather than as a visualization layer added after training.

Related to interpretability is the broader framework of trustworthy AI. Systematic reviews in the recidivism domain repeatedly identify fairness, transparency, privacy and data protection, accountability, human oversight, technical robustness, and social acceptability as core requirements for responsible deployment [10,11,15,29,38,44]. These dimensions matter because a model may achieve acceptable discrimination metrics, still being inappropriate in practice if data provenance is unclear, decision logic is not auditable, or outputs encourage uncritical human reliance. In this sense, interpretability should not be treated merely as a communication aid but rather as a part of the governance structure required for audit, contestability, and responsible human oversight. On the other hand, socio-legal critiques go further by warning that explanation layers attached to opaque systems may convert uncertainty into institutional justification, thereby redistributing responsibility without providing reasons that are scientifically or legally reviewable [43]. For criminal justice applications, this means that technical explanation must be aligned with procedural rights, reviewability, and practical avenues for contesting adverse decisions.

A further concern is transportability and institutional tasks. Studies of race and geography suggest that location can influence predictive performance more consistently than race in some machine learning settings, implying that models developed in one jurisdiction may not generalize reliably to another [25,30,49]. Other work argues that predictive accuracy alone is not enough if the system is meant to support treatment allocation or policy design

rather than only risk ranking; from this perspective, the objective should shift from pure prediction toward learning actionable decision policies [22]. Additional studies on mental illness, intimate partner violence, cognitive-emotion regulation, and juvenile offending show that subgroup-specific recidivism prediction often requires different feature spaces, different operational definitions, and different interpretability settings [16,18,20,26]. These findings suggest that recidivism modeling should be treated as a context-sensitive problem in which fairness, interpretability, and deployment goals must be aligned with the intended institutional task.

Data quality and class imbalance present additional technical barriers as recidivism datasets are often imbalanced, fragmented across institutions, and constrained by privacy or legal restrictions. Recent work addresses these issues through feature-selection pipelines, SMOTE-based balancing, clustering, hyperparameter optimization, synthetic-data generation, and direct AUC-oriented objectives [14,27,40,50]. Open-source replications and synthetic-data studies also highlight the importance of reproducibility, privacy-preserving experimentation, and methodological transparency [35,50]. However, gains in predictive performance do not, by themselves, resolve fairness or interpretability concerns. For example, some studies report high apparent accuracy in specialized or local datasets, whereas others caution that data imbalance, outcome definition, and deployment context can distort naive evaluations. Likewise, work on mental illness suggests that clinically salient variables may add limited predictive value over crime and demographic features, even though they remain important for treatment planning and legal–ethical analysis [26]. These observations reinforce the need for model designs that balance predictive power with transparent reasoning and fairness-sensitive evaluation, rather than optimizing performance in isolation.

The human side of recidivism prediction is equally important. Public attitude studies show that people often underestimate the error rates of algorithmic systems while demanding very low error tolerance in high-stakes contexts [14]. Other work on crowd perceptions and human algorithm interaction indicates that notions of fairness depend not only on statistical metrics but also on certain predictors that are viewed as legitimate, how the system is explained, and whether people retain meaningful oversight [12,36]. In professional settings, algorithmic support can improve human predictions in some groups, especially among targeted or trained users, but practitioners remain reluctant to fully endorse automated recidivism assessment [12]. Instead, such tools are often viewed as aids for standardization, training, cross-checking, and capacity extension. This has direct implications for model design. A fairness-aware and interpretable model is valuable not only because it can be inspected statistically, but also because it can support calibrated, reviewable, and contestable human decision processes.

A review of the literature reveals a clear research gap. Existing studies often optimize one or two properties at a time: some emphasize predictive performance, others focus on fairness auditing, others prioritize explainability, and others examine legal or ethical implications [9,31,33,38,44]. Comparatively fewer studies integrate these objectives into a single design framework for structured recidivism data. Even when fairness and interpretability are both discussed, fairness is frequently treated as a post hoc evaluation layer, while interpretability is treated as a reporting feature rather than as a structural principle embedded in the model itself. Based on the above analysis, the main contributions of the current endeavor are enumerated as follows. First, three recidivism datasets from three European countries were developed. The two of them were provided by official authorities, while the third one was artificially synthesized based on existing statistical distributions and sophisticated data generation methods. Given that structured recidivism datasets are rare in the existing literature, the above contribution provides a significant impact on

studying recidivism and creating machine learning models to quantify its properties and characteristics. Second, it develops a model that remains operationally interpretable. Third, it evaluates the model not only by predictive performance but also by fairness-relevant criteria. Fourth, it contributes to the broader transition from opaque risk scoring toward algorithmic systems that are technically effective, normatively defensible, and institutionally trustworthy. Fifth, it situates the technical design within the wider literature on criminal justice analytics, trustworthy AI, and responsible deployment of AI software, thereby linking model construction to auditability, contestability, and human oversight. In this way, the study is positioned at the intersection of predictive modeling, interpretable machine learning, algorithmic fairness, and criminal justice governance.

The remainder of this paper is organized as follows. Section 2 presents the methodological framework. Section 3 describes the three recidivism datasets acquired from Bulgaria, Greece, and Portugal, detailing their attribute structures, output definitions, and the methodology used to generate the Portuguese dataset. Section 4 presents experimental analysis, organized into several simulation cases. Finally, the paper concludes in Section 5.

## 2. Materials and Methods

### *2.1. Predictive Modeling Framework*

To capture the complex and potentially non-linear relationships between an offender's socio-demographic profile and their likelihood of recidivism, the system employs a deep learning framework. Specifically, the architecture utilizes a 1-Dimensional Convolutional Neural Network (1D-CNN). While traditionally applied to time-series or sequential data, 1D CNNs can be highly effective on tabular data by treating the ordered feature vector as a spatial sequence, thereby extracting localized feature interactions before making a global prediction [48].

The model is structured as a sequential network with the following topological flow: (a) input representation, (b) convolutional feature extraction, and (c) flattening and classification. The functionality of the convolutional layer, denoted as Conv1D, relies on using kernels and feature maps to process one-dimensional tabular input data and extract local patterns (i.e., features). In particular, the feature maps can identify the local dependencies between neighboring input samples. In this regard, the convolution operation uses element-wise multiplication of the feature maps' weights with the corresponding input values and sums them up [48].

On the other hand, the pooling layer is used to progressively reduce the dimensions (width) of the input data as it passes through the network, while retaining the most important information. The more pooling layers there are in a network, the greater the reduction effect. By reducing the spatial dimensions of the input, pooling layers help to decrease the computational complexity of the network. Pooling layers summarize the presence of features within their pooling regions. By retaining the most important features while discarding less relevant details, pooling layers can help extract important features and capture the essence of the input data.

Figure 1 illustrates the basic structure of the ML model used in the current endeavor. Specifically, we used one Conv1D layer followed by a pooling layer and a Conv1D layer. Finally, a flattened dense layer is used before the output layer. In what follows, we provide the mathematical description of the above layers.

Let us assume that we are given an input–output data set of the form $\{X, Y\}$ where $X \in \Re^{n \times m}$ is the input data array and $Y \in \Re^{n}$ the output. Thus, $n$ corresponds to the number of instances (i.e., samples or individuals), and $m$ corresponds to the number of attributes that define the feature space.

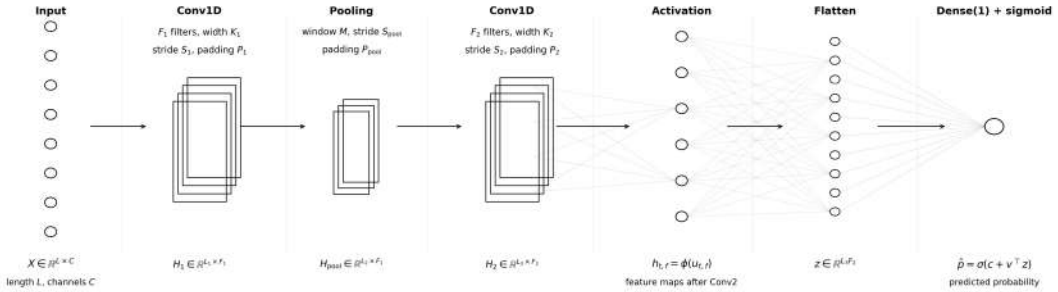


**Figure 1.** The basic structure of the 1D-CNN.

In particular, $X$ can be written as $X = [x_k]|_{k=1}^{n}$ with $x_k$ being equal to $x_k = [x_{k1}, x_{k2}, \ldots, x_{km}]^T \in \Re^m$, while $Y$ can be written as $Y = [y_k]|_{k=1}^{n}$ with $y_k \in \Re$. In our case, we are dealing with binary classification and thus $y_k \in \{0, 1\}\ \forall k$. Based on the above nomenclature, the output vector can also be written as $Y \in \{0, 1\}^n$, and each input sample as $x_k \in \Re^{m \times 1}$.

For the first convolution layer, we define the kernel size as $K_1$, the number of feature maps as $F_1$, the stride parameter as $St_1$, and the padding parameter as $Pd_1$. Thus, the output length is equal to

$$L_1 = \left\lfloor \frac{m - K_1 + 2Pd_1}{St_1} \right\rfloor + 1 \tag{1}$$

and the convolution operation is computed as follows:

$$y_k^{(1,t)}(i) = \sum_{j=0}^{K_1-1} w_k^{(1)}(j)\, x_k^{(t)}(i\, St_1 + j) + b_k^{(1)} \tag{2}$$

where $t = 1, 2, \ldots, n$, $w_k$ are the neural weights and $b_k$ are the respective bias parameters. The activation function is the ReLU function, defined as

$$a_k^{(1,t)}(i) = \max\left\{0,\, y_k^{(1,t)}(i)\right\} \tag{3}$$

The layer's output tensor is

$$a^{(1,t)} \in \Re^{L_1 \times F_1} \tag{4}$$

For the pooling layer, defining the window size as $K_p$ and the stride parameter as $St_p$, the layer's output length is calculated as follows:

$$L_2 = \left\lfloor \frac{L_1 - K_p}{St_p} \right\rfloor + 1 \tag{5}$$

while the max pooling operation is

$$y_k^{(2,t)}(i) = \max_{0 \le j < K_p} \left\{ a_k^{(1,t)}(i\, St_p + j) \right\} \tag{6}$$

and the layer's output tensor operator as

$$y^{(2,t)} \in \Re^{L_2 \times F_1} \tag{7}$$

For the second convolution layer, we denote the kernel size as $K_2$, the number of feature maps as $F_2$, the stride parameter as $St_2$, and the padding parameter as $Pd_2$. As such, the layer's output length is calculated as

$$L_3 = \left\lfloor \frac{L_2 - K_2 + 2\, Pd_2}{St_2} \right\rfloor + 1 \tag{8}$$

Thus, the corresponding convolution operator is given by

$$y_k^{(3,t)}(i) = \sum_{c=1}^{F_1} \sum_{j=0}^{K_2-1} w_{k,c}^{(2)}(m)\, y_c^{(2,t)}(i\,St_2 + j) + b_k^{(2)} \tag{9}$$

Given that the node activation function is the ReLU function,

$$a_k^{(3,t)}(i) = \max\left\{0,\, y_k^{(3,t)}(i)\right\} \tag{10}$$

and the layer's output tensor is defined as

$$a^{(3,t)} \in \Re^{L_3 \times F_2} \tag{11}$$

For the flattened layer, we flatten the above tensor into the subsequent vector:

$$z^{(t)} = vec(a^{(3,t)}) \in R^{L_3 \times F_2} \tag{12}$$

To this end, the operation that defines the whole network's output is calculated as

$$o^{(t)} = w^{(4)}\, z^{(t)} + b^{(4)} \tag{13}$$

To quantify the output probabilities, we use the following sigmoid function:

$$p_t = \frac{1}{1 + \exp(-o^{(t)})} \tag{14}$$

Finally, the objective function that drives the learning process is the binary cross-entropy

$$L_{CE}(w) = -\frac{1}{n} \sum_{t=1}^{n} \left( y^{(t)} \log p_t + (1 - y^{(t)}) \log(1 - p_t) \right) \tag{15}$$

### *2.2. Fairness Analysis*

Fairness in machine learning (ML) has become a central design direction, especially when socially sensitive domains (e.g., criminal justice, hiring, financial sector, institution entry, etc.) are involved [7]. Trained ML models on historical data often inherit and grow prejudices that existed in the data, leading to discriminatory decision-making against certain demographic groups [2,4]. A group that is favored by the ML-model decisions is called privileged or non-sensitive, whereas in the opposite case, it is called unprivileged or sensitive. The privileged and unprivileged groups are defined by a binary attribute called a protected attribute. To anticipate such kind of discrimination and counterbalance its effects, fairness criteria are involved in the model's development process with the ultimate purpose of mitigating the corresponding bias that exists in the training data and providing fairer ML-based decision-making regarding the unprivileged group [3,28]. Most current measures of fairness, such as Equalized Odds (EO), demographic parity (DP), or Predictive Parity (PP), purportedly consider discrete sensitive characteristics like gender, race, etc. In particular, the EO criterion has been widely involved in a fair ML perspective because it requires simultaneous minimization of Type I and Type II errors [2,3]. Minimization of Type I error implies that the false positive rates (FPRs) between the privileged and unprivileged groups must be similar. Relationally, minimization of Type II error assumes that the false negative rates (FNRs) must be similar across the above-mentioned groups.

Herein, the basic mathematical formulation of the EO fairness metric is described. Let us assume that the available data set consists of $n$ input–output data, where the input data are defined in an $m$-dimensional feature space, and the discrete random variable $S$ that corresponds to the protected attribute,

$$D = \{X, S, Y\} \tag{16}$$

where $X = [x_k]|_{k=1}^{n}$ is the input feature matrix, input attributes with $x_k \in \Re^m$, $Y = [y_k]|_{k=1}^{n}$ is the output attribute with $y_k \in \{0, 1\}$, and $S = [s_k]|_{k=1}^{n}$ with $s_k$ being equal to 1 if $x_k$ belongs to the privileged group otherwise, it is equal to 0. For the random variables $S$ and $Y$, we can also write $S \in \{0, 1\}$ and $Y \in \{0, 1\}$.

**Definition 1 (Equalized Odds).** *Given the data set $D = \{X,\ S,\ Y\}$ and the classifier's predicted output $\hat{Y} \in \{0,\ 1\}$, we say that the classifier satisfies the Equalized Odds (EO) criterion with respect to $S$ and $Y$ if $\hat{Y}$ and $S$ are conditionally independent given $Y$, which implies that, $\forall y \in \{0, 1\}$ it holds that*

$$P\big(\hat{Y} = 1 | Y = y,\ S = 0\big) = P\big(\hat{Y} = 1 | Y = y,\ S = 1\big) \tag{17}$$

Note that Equation (17) can be decomposed into the following two equations:

$$P\big(\hat{Y} = 1 | Y = 1,\ S = 0\big) = P\big(\hat{Y} = 1 | Y = 1,\ S = 1\big) \tag{18}$$

$$P\big(\hat{Y} = 1 | Y = 0,\ S = 0\big) = P\big(\hat{Y} = 1 | Y = 0,\ S = 1\big) \tag{19}$$

Equation (18) states that the true positive rates (TPRs) across the two groups defined by the protected attribute should be equal. Relationally, Equation (19) postulates that the FPRs across the above groups should also be equal. Regarding Equation (18), whenever the TPRs are equal, the same holds for the FNRs, a fact described by the subsequent modification of Equation (18),

$$P\big(\hat{Y} = 0 | Y = 1,\ S = 0\big) = P\big(\hat{Y} = 0 | Y = 1,\ S = 1\big) \tag{20}$$

We say that the classifier $\hat{Y}$ fulfills the EO criterion if the conditions (19) and (20) are simultaneously satisfied. FPRs and FNRs are also known as Type I and II error rates, respectively [1–3]. Both refer to the probability equality of a person in the unprivileged group being assigned a positive outcome [29]. The importance of Equations (19) and (20) is evident in areas where decisions are affected by the results of a model. Equal FPRs and FNRs provide significant potential to the model in correctly predicting positive cases. However, when FNRs or FPRs are disproportionate, the model's behavior changes across groups. Especially in the case of FNRs, that situation might yield large Type II error rates, because high disparity in FNRs implies that instances from different groups have been incorrectly classified. When strong imbalances between the two parts of Equation (19) are present, higher FNR values correspond to the privileged group, whereas lower FNR values are related to stricter predictions of the classifier and correspond to the unprivileged group [44]. Similar results can be identified in the case of FPRs (i.e., Equation (20)) that concerns the Type I error rates.

### *2.3. Model's Optimization Process*

The task studied in this section concerns the bias mitigation process, where the bias is quantified in terms of the Equalized Odds (EO) criterion. To accomplish that task, we develop a custom loss (i.e., objective) function that inserts the EO criterion as a constraint into the binary cross-entropy given in Equation (15). Thus, the training process of the

machine learning model is based on minimizing that loss function instead of the pure cross-entropy. As such, the proposed loss function consists of the standard binary cross-entropy focusing on optimizing the network regarding the classification accuracy, and a fairness-specific part that constrains the above optimization procedure by quantifying the EO criterion. The EO loss penalizes the model for unequal FPRs and FNRs across groups defined by the protected attribute, promoting balanced treatment. The resulting constraints are considered in the optimization scheme as an L1-regularization (i.e., LASSO) approach, which gradually increases the contribution of the fairness term over training epochs to avoid early destabilization of the minimization procedure. Given that the probabilities as calculated by Equation (14) are, $p_i = p_i(w)$ $(i = 1, 2, \ldots, n)$, the binary cross-entropy is rewritten as indicated next

$$L_{CE}(w) = -\frac{1}{n}\sum_{i=1}^{n}(y_i \ln(p_i) + (1 - y_i)\ln(1 - p_i)) \tag{21}$$

The FPR condition in Equation (19) is related to the following function [3,51]:

$$h_{FPR}(w) = \left| \frac{\sum_i p_i(1 - y_i)s_i}{\sum_i s_i} - \frac{\sum_i p_i(1 - y_i)(1 - s_i)}{\sum_i (1 - s_i)} \right| \tag{22}$$

Relationally, the FNR condition related to Equation (20) can be written as:

$$h_{FNR}(w) = \left| \frac{\sum_i (1 - p_i)\, y_i\, s_i}{\sum_i s_i} - \frac{\sum_i (1 - p_i)\, y_i\, (1 - s_i)}{\sum_i (1 - s_i)} \right| \tag{23}$$

Thus, the constrained optimization problem imposed by the EO criterion, as described in Equations (22) and (23), reads as follows:

$$\begin{aligned} &\text{minimize} \quad L_{CE}(w) \\ &\text{subject to} \quad H_{EO1}(w) = (h_{FPR}(w))^2 - \delta \leq 0 \\ &\qquad\qquad\quad H_{EO2}(w) = (h_{FNR}(w))^2 - \delta \leq 0 \end{aligned} \tag{24}$$

Herein, we resolve the above problem using the following regularization approach:

$$L(w) = L_{CE}(w) + \alpha\,(H_{EO1}(w) + H_{EO2}(w)) \tag{25}$$

where $\alpha$ is the regularization parameter that controls the counterbalance between the two parts in Equation (25). The total loss function $L(w)$ is minimized in terms of the stochastic gradient descent optimizer. As shown in Equation (25), the two parts are the loss function $L_{CE}(w)$ and the fairness loss function $H_{EO}(w) = H_{EO1}(w) + H_{EO2}(w)$. That customizes the prediction accuracy and fairness, where inequalities in mistake rates among privileged and unprivileged groups are obviously penalized by the model's inclusion of fairness loss in its aim. This method guarantees that the model will continue to perform similarly across a range of those two groups once it has learned to accurately identify instances.

### *2.4. Model Interpretability*

To provide case-level interpretability of the classifier, we employed the Kernel SHAP (SHapley Additive exPlanations) approach, which constitutes a model-agnostic explanation framework derived from Shapley value theory [41,46]. The central idea is to decompose the prediction for a given instance into an expected model output plus additive contributions from the individual input features. Because most machine learning models

cannot be evaluated directly on arbitrary subsets of observed and missing variables, Kernel SHAP approximates these contributions by masking features with respect to a background data set and fitting a weighted local surrogate model whose coefficients correspond to Shapley attributions. As a result, the method yields a mathematically principled explanation of how each feature contributes to the deviation of the prediction from its baseline value, making it well-suited for the interpretation of individual risk predictions in applied classification settings.

For a given instance $x$, Kernel SHAP explains the model prediction $f(x)$ with an additive surrogate model of the form:

$$g(z) = \varphi_0 + \sum_{i=1}^{M} \varphi_i z_i \tag{26}$$

where $z \in \{0,1\}^M$ is a binary coalition vector over the $M$ interpretable features, $\varphi_0$ is the baseline prediction, and $\varphi_i$ is the attribution assigned to the feature $i$. In the SHAP framework, these coefficients are chosen so that the explanation is locally accurate and additive, that is, the prediction is decomposed as the baseline plus the sum of the feature contributions. The theoretical basis of Kernel SHAP is the Shapley value from cooperative game theory. For feature $i$, the Shapley attribution can be written as follows:

$$\varphi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} [u_x(S \cup \{i\}) - u_x(S)] \tag{27}$$

where $N = \{1, \ldots, M\}$ is the full feature set and $u_x(S)$ denotes the model value associated with the coalition $S$. This expression averages the marginal contribution of the feature $i$ over all possible coalitions, with combinatorial weights that ensure a fair allocation of the prediction among the input features.

Because the predictive model cannot directly evaluate arbitrary subsets of observed and missing features, we define coalition values using a background data set. In practice, features outside the coalition are replaced with values drawn from the background data, so that "missingness" is simulated rather than passed literally to the model. This formulation is as follows:

$$u_x(S) = \mathbb{E}_{b \sim B}\left[f\left(x_S, b_{\overline{S}}\right)\right] \tag{28}$$

where $B$ is the background distribution, $x_S$ are the observed feature values retained from the explained instance, and $b_S$ are the complementary values supplied by the background sample. Consequently, both the baseline $\varphi_0$ and the feature attributions $\varphi_i$ depend on the choice of the background data set.

Kernel SHAP estimates the Shapley values by fitting the additive surrogate model through a weighted least-squares problem over the following sampled coalitions:

$$\min_{\varphi_0, \varphi_1, \ldots, \varphi_M} \sum_{z} \pi_x(z) \left[ f(h_x(z)) - \left( \varphi_0 + \sum_{i=1}^{M} \varphi_i z_i \right) \right] \tag{29}$$

where $h_x(z)$ maps the binary coalition vector into a masked input instance and $\pi_x(z)$ is the SHAP kernel. The kernel used in Kernel SHAP is

$$\pi_x(z) = \frac{M - 1}{\binom{M}{|z|} |z|(M - |z|)} \quad 0 < |z| < M \tag{30}$$

which gives higher weight to very small and very large coalitions and yields the Shapley value solution within the additive explanation class.

The resulting explanation satisfies the following local additive decomposition

$$f(x) \approx \varphi_0 + \sum_{i=1}^{M} \varphi_i \tag{31}$$

The explainer uses a logit link, and the attributions become additive in log-odds space

$$\log it(f(x)) \approx \varphi_0 + \sum_{i=1}^{M} \varphi_i \tag{32}$$

## 3. Description of the Datasets

In this section, we present three new recidivism datasets acquired in the current endeavor from three different countries, namely, Bulgaria, Greece, and Portugal. The Bulgarian and Greek datasets were, respectively, taken from the Bulgarian and Greek official authorities and were processed to be appropriately anonymized, fully compliant with the EU General Data Protection Regulation (GDPR). For the Portuguese case, no official datasets were available. Therefore, that data set was artificially generated by using publicly available statistics and the Greek data set. The generation process was based on using the data distributions coming from the public Portuguese statistics and the respective Greek distributions. The final variables, data instances, and simulation outcomes for the Portuguese data set were thoroughly studied and checked by a group of experts, such as lawyers, establishment officers, judges, etc., a fact that guarantees both realism and generalizability of the generated data and the experimental findings reported in the next section (i.e., see Section 4).

Herein, recidivism has been specifically defined as reincarceration following release from prison. This definition excludes rearrests that do not result in custody, technical probation violations (unless leading to custodial admission), and purely administrative returns. The operationalization differs slightly across national contexts due to the data set architecture. Regarding the Bulgarian data set, due to data constraints, an event-based proxy measure is constructed by the authorities, which distinguishes custodial admission from probation admission. In the Greek data set, recidivism has been defined as reincarceration due to a new offense within a defined post-release time window (typically within three years). Finally, in the Portuguese data set, recidivism risk is estimated through structured administrative variables aligned with validated criminological predictors. These differences reflect responsible adaptation to national institutional realities rather than conceptual divergence.

Tables 1–3 depict the attribute names and types for the Bulgarian, Greek, and Portuguese datasets, respectively. There are 4940, 12,422, and 4418 sample instances available for the Bulgarian, Greek, and Portuguese datasets, respectively. Each sample instance corresponds to an individual.

The Bulgarian data set includes 17 input attributes and one output attribute, called "Recidivism Risk Assessment". It is worth noting that the output attribute, while not straightforwardly referring to re-incarceration as a binary-typed variable, quantifies the re-incarceration risk assessment carried out by the Bulgarian official authorities as low and medium.

The Greek data set includes 11 input variables and two outputs. The two outputs are called "Recidivism within 3 Years" and "Recidivism". The first one quantifies the probability of reincarceration in a three-year interval after an individual's prison release, while the second one quantifies the probability of reincarceration in an individual's life span after prison release. Note that the first output is widely considered a reliable index to effectively predict the recidivism related to an individual's attribute status.

**Table 1.** Description of the attributes for the Bulgarian data set.

| Attribute Name | Attribute Type and Values | Attribute Name | Attribute Type and Values | Attribute Name | Attribute Type and Values |
|---|---|---|---|---|---|
| Gender | Binary {male, female} | Siblings | Binary {yes, no} | Revocation of the release (service of an unserved sentence) | Binary {yes, no} |
| Nationality | Binary {Bulgarian, foreigner} | Prison status | Binary {measure detained, penalty sentenced} | Sentence fulfilled | Binary {yes, no} |
| Age at entering prison | Real | Days in prison | Real | Interruption of implementation | Binary {yes, no} |
| Age at exiting prison | Real | Days of penalty | Real | Conditional early release | Binary {yes, no} |
| Level of education | Categorical * {illiterate, basic, primary, secondary, high school, higher education} | Sentence for multiple crimes | Binary {yes, no} | Reducing punishment with work | Binary {yes, no} |
| Marital status | Categorical * {single, divorced, married, other} | Exemption from serving the sentence | Binary {yes, no} | Recidivism risk assessment ** | Binary {low, high} |

* Transformed into one-hot encoding; ** Output attribute.

**Table 2.** Description of the attributes for the Greek data set.

| Attribute Name | Attribute Type and Values | Attribute Name | Attribute Type and Values | Attribute Name | Attribute Type and Values |
|---|---|---|---|---|---|
| Gender | Binary {male, female} | Age of release | Real | Crime category | Categorical (it contains 11 crime categories) |
| Nationality | Binary {Greek, foreigner} | Penal situation | Categorical {conviction, pre-trial detention, imprisonment} | Recidivism within 3 years ** | Binary {yes, no} |
| Level of education * | Categorical {illiterate, basic, lower secondary school, higher secondary school, higher education, other} | Reason for release * | Categorical [1] | Recidivism ** | Binary {yes, no} |
| Family status * | Categorical {single, married, cohabitation, divorced, other} | Days in prison | Real | | |
| Employment * | Categorical {employed, unemployed, other} | Sentence length | Real | | |

* Transformed into one-hot encoding; ** Output variables; [1] {conditional release, end of sentence, installment payment of the fine, lodging an appeal, release from custody, sentence conversion, substitution of pre-trial detention, suspension, other}.

**Table 3.** Description of the attributes for the Portuguese data set.

| Attribute Name | Attribute Type and Values | Attribute Name | Attribute Type and Values | Attribute Name | Attribute Type and Values |
|---|---|---|---|---|---|
| Gender | Binary {male, female} | Sentence length * | Categorical {Up to 12 months, 1 to 3 years, 3 to 6 years, 6 to 9 years, 9 to 15 years, 15 to 25 years} | Reason for release * | Categorical {conditional release, end of pre-trial detention, end of the sentence, other reasons} |
| Nationality | Binary {Portuguese, foreigner} | Crime category * | Categorical {crimes against persons, crimes against property, crimes against society, crimes against the State, drug-related crimes, other crimes} | Recidivism within 3 years ** | Binary {yes, no} |
| Age of release | Categorical {[16, 18], [19, 20], [21, 24], [25, 29], [30, 39], [40, 49], [50, 59]}, over 60 | Type of crime | Categorical (it contains 15 types of crime) | Recidivism ** | Binary {yes, no} |
| Education level * | Categorical {illiterate, 1st basic (1–4 years), 2nd basic (5–6 years), 3rd basic (7–9 years), secondary (10–12 years, higher education} | Management complexity | Binary {medium, high} | | |

* Transformed into one-hot encoding; ** Output variables.

The Portuguese data set includes nine input attributes and two output attributes. The output attributes are similar to the Greek data case.

In view of the above tables, we can distinguish between two types of attributes, namely, static and dynamic attributes. Static attributes refer to factors that are fixed or not meaningfully modifiable in the short term, such as gender, nationality, age at exiting prison, criminal patterns, structural offense characteristics, etc. Static factors consistently demonstrate a strong statistical association with recidivism. Age, for instance, follows the well-established "age–crime curve," where criminal involvement peaks in late adolescence and early adulthood and declines with age. Criminal patterns, such as the type of crime or crime category, are among the most powerful predictors of future criminal justice contact. In general, the advantage of static variables lies in their objectivity and reproducibility. They reduce subjective interpretation and are reliably documented in administrative datasets. On the other hand, dynamic attributes correspond to dynamic factors able to change over time and may be influenced by intervention. Examples include employment status, educational status, institutional behavior, etc. Dynamic factors are highly relevant for rehabilitation-oriented policy. However, they are often inconsistently recorded across national administrative systems. For this reason, the current endeavor relies primarily on static predictors supplemented by limited dynamic factors for socio-economic vulnerability.

Regarding the output variables, in the Bulgarian case, no information was provided to set up a model where the output would be the individual's risk assessment related to the probability of reincarceration in a three-year time interval after his/her prison release. As far as the other two cases are concerned, the data structure is defined in a straightforward manner in terms of the two output attributes, namely "Recidivism" and "Recidivism within 3 Years". Therefore, for the Bulgarian case, only one classification model is constructed to predict the re-incarceration risk, while for each one of the Greek and Portuguese cases,

two models will be constructed. The first model concerns the "Recidivism within 3 years" output, and the second one the "Recidivism" output.

## 4. Experiments and Discussion

Herein, we apply the methodology proposed in Section 2 on the datasets described in Section 3 to estimate and mitigate bias against gender. Thus, the sensitive attribute is the gender attribute. This attribute is divided into the Male group, defining the privileged group, and the Female group, defining the unprivileged group.

Recalling what was reported in the previous section, there are three available datasets for Bulgaria, Greece, and Portugal. The Bulgarian data set has only one binary output that corresponds to the re-incarceration risk assessment and is called "Recidivism" or "Recidivism risk assessment". The other two datasets include two outputs called "Recidivism" and "Recidivism within 3 Years", where the first one quantifies the probability of reincarceration in a three-year interval after an individual's prison release, and the second quantifies the probability of reincarceration in an individual's life span after prison releaseThe objectives of the experimental analysis are enumerated as follows: (a) study of the model's accuracy before and after the bias mitigation, (b) study of the bias against gender before and after the mitigation process based on EO fairness criterion, (c) study of the Predictive Parity (PP) fairness criterion, and (d) study of the interpretability capabilities of the model after the mitigation process, as far the EO criterion is concerned.

To evaluate the behavior without bias mitigation, we build the 1D-CNN model using the cross-entropy objective function defined in Equations (15) and (21). Therefore, this model favors the bias estimation process. On the other hand, we use the custom objective function given in (25) to create a model that performs bias mitigation and study the model's behavior after bias mitigation.

In view of Tables 1–3, the created 1D-CNN models are subsequently described. First, for the Bulgarian data set, we built two models to study the "Recidivism risk assessment" output attribute, before and after the mitigation process, respectively. Second, for the Greek data set, we built two models to study the "Recidivism within 3 years" output attribute, before and after the mitigation process, respectively. Third, for the Greek data set, we built two models to study the "Recidivism" output attribute, before and after the mitigation process, respectively. Fourth, for the Portuguese data set, we built two models to study the "Recidivism within 3 years" output attribute, before and after the mitigation process, respectively. And finally, for the Portuguese data set, we built two models to study the "Recidivism" output attribute, before and after the mitigation process, respectively. Thus, in total, we create five models to quantify the behaviors before bias mitigation and five models to evaluate the status after bias mitigation.

To train the 1D-CNN models, we used the stochastic gradient descent (SGD) algorithm, where the activation functions were quantified in terms of the ReLU function, while the learning rate, the number of maximum epochs, and the batch size were set to 0.0001, 2000, and 50, respectively. The parameters for 1D-CNN were as follows: number of kernels = 16, kernel size = 2, padding = "same", input_shape = ($m$, 1), where $m$ is the number of input attributes, and input channels = 1. Thus, the total number of parameters was equal to 48 (i.e., 32 weights and 18 biases). The regularization parameter $\alpha$ in Equation (25) was determined in terms of an iterative process where its initial value was very small (i.e., favoring the presence of bias), and as the iteration number increased, this value also increased using a pre-defined step size. The iteration stops when, during two consecutive iterations, the minimization rates of the fairness part in Equation (25) are close to each other. The final values for the Bulgarian, Greek, and Portuguese datasets were $\alpha$ = 10, 5, and 7, respectively. Finally, for each simulation case, the original data set

was divided into a training set, including the 70% of the data, and a testing set, including the rest 30% of the data.

To carry out the statistical experimental analysis, considering all 10 1D-CNN models, 100 runs for each model were executed using different initializations. The results reported in the following subsections concern the testing data.

Based on the above setting, we performed four experimental cases, called Experiment 1, Experiment 2, Experiment 3, and Experiment 4, which are presented within the next subsections.

### *4.1. Experiment 1: Descriptive Statistics of the Accuracy Performance*

This case concerns the study of the model's accuracy before and after the mitigation process. As a first step, apart from testing the 1D-CNN model, we also tested two more models, namely, the standard XGBoost algorithm and a standard MLP neural network. The MLP consisted of two hidden layers with 10 and 5 nodes, respectively, and the ReLU function as the nodes' activation operator. Table 4 depicts the simulation results in terms of the accuracy obtained over the testing data.

**Table 4.** Accuracies obtained by XGBoost and MLP models (100 runs for simulation case) on the testing data.

| | | **Recidivism Prediction** | | | **Within 3 Years Recidivism Prediction** | |
|---|---|---|---|---|---|---|
| | | **Bulgarian** | **Greek** | **Portuguese** | **Greek** | **Portuguese** |
| XGBoost | Mean | 0.7056 | 0.6998 | 0.8624 | 0.7237 | 0.9107 |
| | Std | 0.0081 | 0.0129 | 0.0059 | 0.0139 | 0.0056 |
| MLP | Mean | 0.7321 | 0.6922 | 0.8389 | 0.7279 | 0.8715 |
| | Std | 0.0108 | 0.0076 | 0.0083 | 0.0078 | 0.0076 |

Regarding the implementation of the 1D-CNN model, the results are depicted in Tables 5 and 6.

**Table 5.** Accuracy obtained by the 1D-CNN models before and after the bias mitigation process for the recidivism prediction case (100 runs for each model) on the testing data.

| | **Before Bias Mitigation** | | | **After Bias Mitigation** | | |
|---|---|---|---|---|---|---|
| | **Bulgarian** | **Greek** | **Portuguese** | **Bulgarian** | **Greek** | **Portuguese** |
| Mean | 0.7553 | 0.7009 | 0.8626 | 0.6792 | 0.6706 | 0.7845 |
| Std | 0.0097 | 0.0064 | 0.0087 | 0.0285 | 0.0124 | 0.0246 |

**Table 6.** Accuracy obtained by the 1D-CNN models before and after the bias mitigation process for the within 3 years recidivism prediction case (100 runs for each model) on the testing data.

| | **Before Bias Mitigation** | | **After Bias Mitigation** | |
|---|---|---|---|---|
| | **Greek** | **Portuguese** | **Greek** | **Portuguese** |
| Mean | 0.7346 | 0.9106 | 0.6886 | 0.8473 |
| Std | 0.0059 | 0.0065 | 0.0139 | 0.0275 |

Based on the results reported in the tables above, we justify our choice to employ the 1D-CNN in the design process of our algorithmic architecture. The first reason relies on its accurate performance. By comparing the results reported in Table 4 with those reported in Tables 5 and 6 for the case "Before Bias Mitigation", we can easily observe

that, apart from the case of Portuguese data in the "Within 3 Years Recidivism Prediction" case, where the XGBoost slightly outperformed the 1D-CNN model, the 1D-CNN clearly obtained better performance than XGBoost and MLP models. It turns out that the use of convolutional kernels seems to be appropriate for describing and quantifying criminal variables. The second reason relies on the choice of using the Kernel SHAP algorithm to carry out interpretability analysis. In this regard, the Kernel SHAP fits better with the inherent layered structure of the 1D-CNN than the respective algorithmic structures of the XGBoost and MLP models.

Next, we proceed to studying Tables 5 and 6, which exclusively refer to the proposed algorithmic structure. Recalling that, contrary to the implementation of Equation (21), Equation (25) includes fairness constraints, the following conclusions can be extracted. First, the results indicate similar behavioral trends across the models. Second, the best performance is achieved by the Portuguese data set, and the worst by the Greek data set. Third, it is obvious that the implementation of the mitigation process in terms of the customized objective function in Equation (25) compromises the accuracy. Thus, in all experimental cases, the accuracy drops for the debiased models. This reduction was expected in the first place, because the constraints that enable fair behavior are considered in Equation (25), imposing conflicting effects as far as the accuracy is concerned. A more rigorous analysis of this issue is presented at the end of Section 4.2.

### *4.2. Experiment 2: Inference Statistics for Bias Estimation and Mitigation*

In this experiment, we evaluate the descriptive statistics of the 10 1D-CNN models (i.e., five before bias mitigation, and five after bias mitigation). Figures 2–6 illustrate the results. These figures include the following information: (a) the FPR for Male and Female groups and the corresponding differences between the two FPRs, called DFPR (difference in FPR), before and after the bias mitigation process, and (b) the FNR for Male and Female groups and the corresponding differences between the two FNRs, called DFNR (difference in FNR), before and after the bias mitigation process. Note that DFPR and DFNR are quantified by Equations (22) and (23), respectively. For further analysis and discussion, we recall that figures labeled as "Before Bias Mitigation" refer to the bias estimation process, while figures labeled as "After Bias Mitigation" correspond to the bias mitigation process.

From these figures, it is clearly observed that the likelihood of predicting recidivism between Male and Female groups appears to have similar behavior in all simulations that correspond to the "Before Bias Mitigation" case. Similar conclusions are extracted regarding the "After Bias Mitigation" case.

Next, we proceed to studying the cases before and after bias mitigation.

First, we study the model predictions without considering bias mitigation (i.e., before the bias mitigation process). Thus, the main purpose is to perform bias estimation and provide rigorous statistical evidence.

As illustrated in Figures 2a,c, 3a,c, 4a,c, 5a,c and 6a,c, we can extract the following remarks:

- *Remark 1*: In all datasets, the FPR mean values for the Female group are larger than the FPR mean values for the Male group, i.e., FPR_Mean (Male) < FPR_Mean (Female).
- *Remark* 2: For the Greek and Portuguese datasets, the FNR mean values for the Female group are smaller than the FNR mean values for the Male group, i.e., FNR_Mean (Male) > FNR_Mean (Female).
- *Remark* 3: For the Bulgarian data set, the FNR mean values for the Female group are larger than the FNR mean values for the Male group, i.e., FNR_Mean (Male) < FNR_Mean (Female).

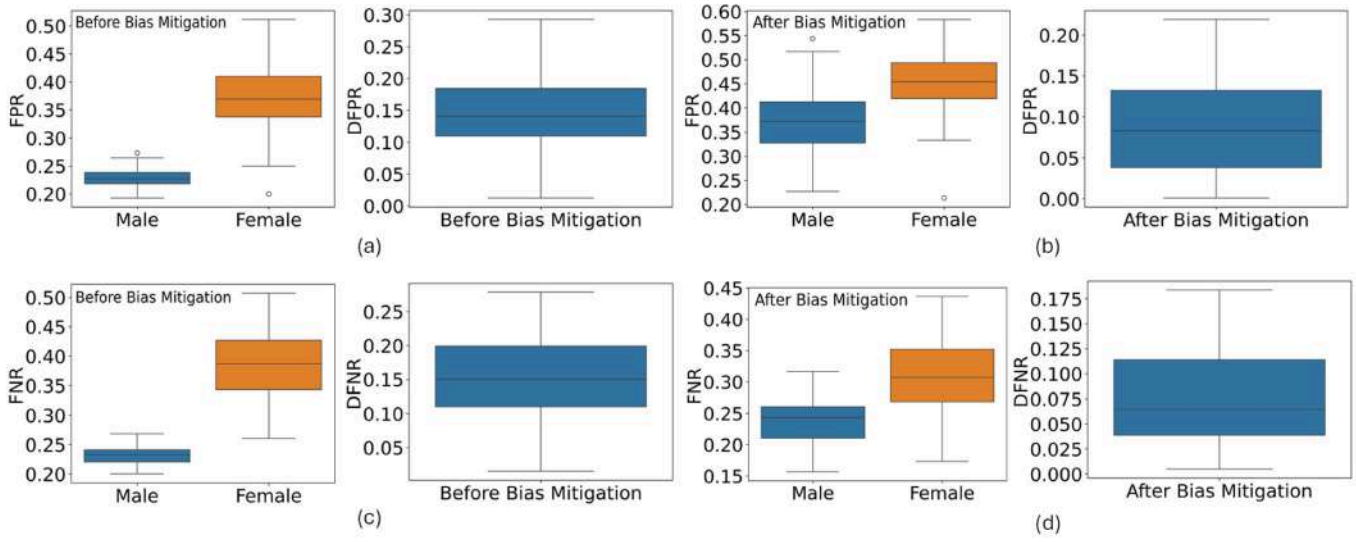


**Figure 2.** Bulgarian data set: Box-plots and the 95% confidence intervals for "Recidivism risk assessment" prediction case (100 runs for each model). (**a**) FPR and DFPR without bias mitigation, (**b**) FPR and DFPR with bias mitigation, (**c**) FNR and DFNR without bias mitigation, (**d**) FNR and DFNR with bias mitigation.

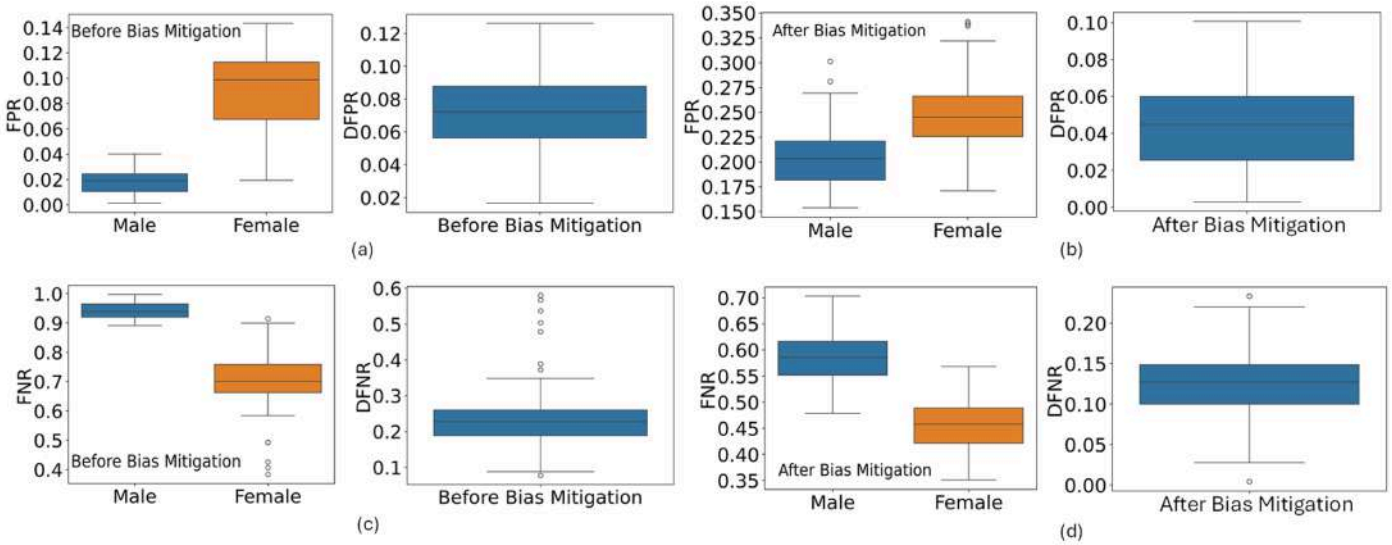


**Figure 3.** Greek data set: Box-plots and the 95% confidence intervals for "Recidivism within three 3 years" prediction case (100 runs for each model). (**a**) FPR and DFPR without bias mitigation, (**b**) FPR and DFPR with bias mitigation, (**c**) FNR and DFNR without bias mitigation, (**d**) FNR and DFNR with bias mitigation.

Next, we carry out rigorous statistical inference to study the distributions reported in Figures 2a,c, 3a,c, 4a,c, 5a,c and 6a,c by considering only FPR and FNR values (i.e., we do not consider the DFPR and DFNR values).

To perform the normality check for those distributions, we employed the well-known Shapiro–Wilk test, with the following Null Hypothesis: "*The population follows normal distribution*". Table 7 depicts the obtained results, where pairs of Male/Female distributions are reported. The reason is that the inference statistics that follow take place considering these pairs of distributions.

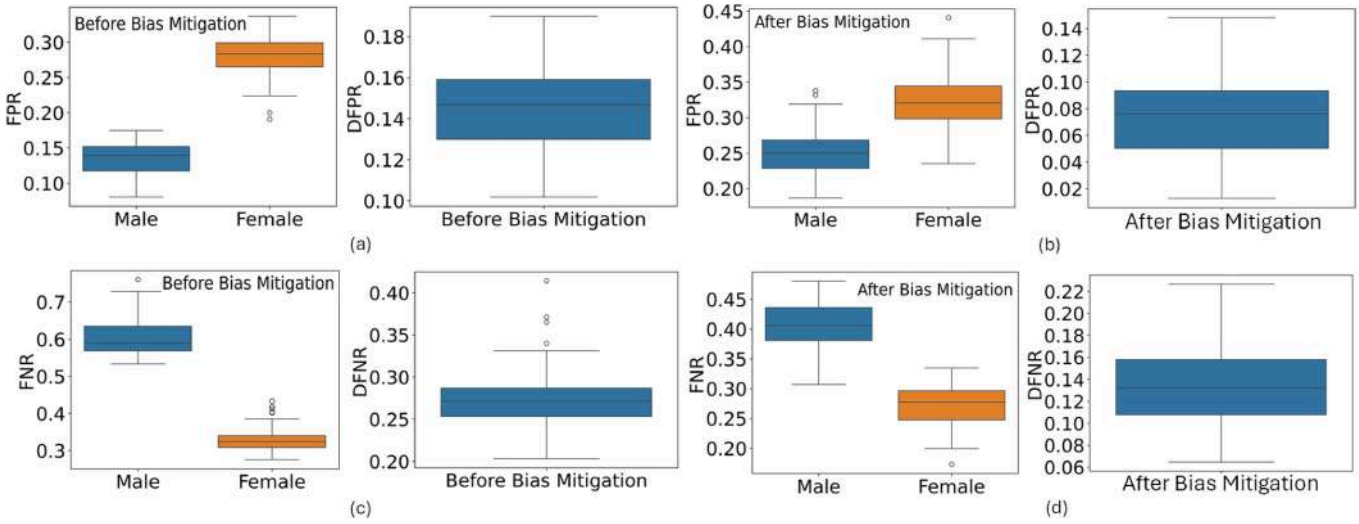


**Figure 4.** Greek data set: Box-plots and the 95% confidence intervals for "Recidivism" prediction case (100 runs for each model). (**a**) FPR and DFPR without bias mitigation, (**b**) FPR and DFPR with bias mitigation, (**c**) FNR and DFNR without bias mitigation, (**d**) FNR and DFNR with bias mitigation.

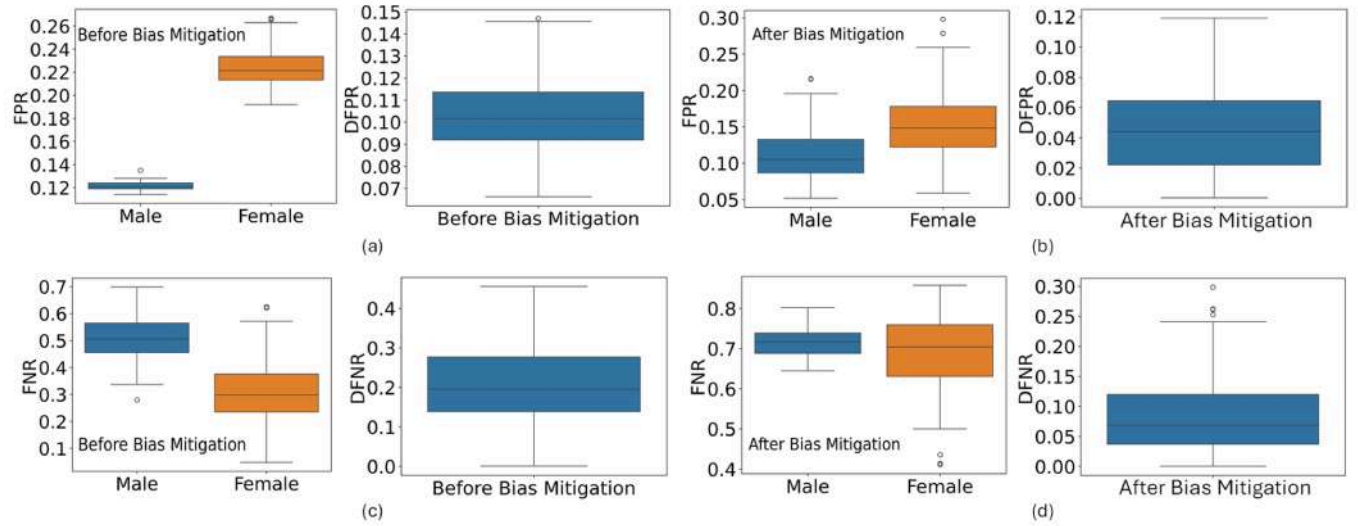


**Figure 5.** Portuguese data set: Box-plots and the 95% confidence intervals for "Recidivism within three 3 years" prediction case (100 runs for each model). (**a**) FPR and DFPR without bias mitigation, (**b**) FPR and DFPR with bias mitigation, (**c**) FNR and DFNR without bias mitigation, (**d**) FNR and DFNR with bias mitigation.

**Table 7.** Acceptance/rejection of the null hypothesis obtained by the Shapiro–Wilk normality check test for models created without bias mitigation.

| Data Set/Output Attribute | FPR for Males/Females | FNR for Males/Females |
|---|---|---|
| Bulgarian/Recidivism risk assessment | Accept | Accept |
| Greek/Recidivism within three 3 years | Reject | Reject |
| Greek/Recidivism | Accept | Reject |
| Portuguese/Recidivism within three 3 years | Reject | Accept |
| Portuguese/Recidivism | Reject | Accept |

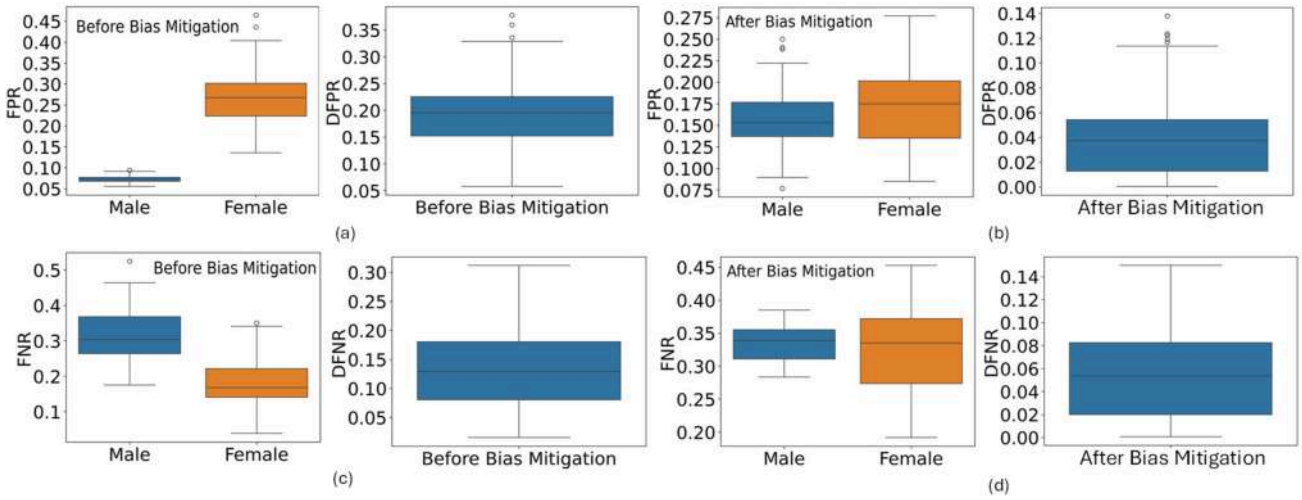


**Figure 6.** Portuguese data set: Box-plots and the 95% confidence intervals for "Recidivism" prediction case (100 runs for each model). (**a**) FPR and DFPR without bias mitigation, (**b**) FPR and DFPR with bias mitigation, (**c**) FNR and DFNR without bias mitigation, (**d**) FNR and DFNR with bias mitigation.

Having said that, in view of Table 7, the inference statistics were carried out in terms of the *t*-test for the cases where the null hypothesis is accepted and the Mann–Whitney U test for the cases where the null hypothesis is rejected. In all comparative cases, the Null Hypothesis is as follows: "*The two populations, corresponding to Male and Female groups, have the same central tendency, which is interpreted as equal distributions*".

Table 8 summarizes the findings of our analysis. The results in these tables directly indicate that the obtained *p*-values are less than 0.05 and therefore the above null hypothesis is rejected in all simulation cases.

**Table 8.** Inference statistics results based on *t*-test (acceptance case in Table 6) or Mann–Whitney U (rejection case in Table 6) for models created without bias mitigation.

| Data Set/Output Attribute | FPR for Males/Females | | FNR for Males/Females | |
|---|---|---|---|---|
| | Statistical Metric | *p*-Value | Statistical Metric | *p*-Value |
| Bulgarian/Recidivism risk assessment | −26.6392 | $1.05 \times 10^{-34}$ | −26.5841 | $1.45 \times 10^{-23}$ |
| Greek/Recidivism within three 3 years | 95 | $4.33 \times 10^{-33}$ | 9972 | $5.42 \times 10^{-34}$ |
| Greek/Recidivism | −41.034 | $9.2 \times 10^{-99}$ | 10,000 | $2.56 \times 10^{-34}$ |
| Portuguese/Recidivism within three 3 years | 0 | $2.56 \times 10^{-34}$ | 14.55675 | $4.07 \times 10^{-33}$ |
| Portuguese/Recidivism | 0 | $3.23 \times 10^{-23}$ | 14.051837 | $4.44 \times 10^{-32}$ |

This means that the populations of the Male group and Female group for all simulation cases regarding FPRs and FNRS are different. That outcome strongly supports the previous analysis that concerned the existence of bias against the Female group.

Trying to analyze the above results, we proceed with the subsequent analysis.

FPR is defined as the ratio between the number of individuals who are predicted to reoffend, while they do not do so, divided by the number of all individuals who do not reoffend. As such, a classifier discriminates against a specific group of individuals when assigning higher FPR values for that group. Thus, in all datasets and in all simulations depicted in Figures 2a, 3a, 4a, 5a and 6a and Tables 7 and 8, the resulting classification models appear to discriminate against the Female group.

FNR is defined as the ratio between the number of individuals who are predicted not to reoffend while they do so, divided by the total number of individuals who reoffend. This

means that a classifier discriminates against a specific group of individuals when assigning it lower FNR values. Thus, in Greek and Portuguese datasets and in all simulations depicted in Figures 3c, 4c, 5c and 6c and Tables 7 and 8, the resulting classification models appear to discriminate against the female group.

Based on the above discussion, Definition 1 and Equations (17)–(20), we can easily verify that for the Greek and Portuguese datasets, the developed classification models exhibit discriminative behavior against the Female group by violating the Equalized Odds requirement. This directly implies that the machine learning models that do not perform bias mitigation appear to have a strong bias against the unprivileged group.

As far as the Bulgarian data set is concerned, while the results for the FPR show clear discrimination against the Female group, this outcome is not clear for the FNR case. However, as will be shown, also in this case, the models appear to exhibit discriminative behavior against the Female group.

To interpret those conclusions, we adopt the concept of base rate (BR) and proceed with the following analysis. The BR is defined as the proportion of individuals in the population who reoffend. In all datasets, women tend to have lower BRs of reoffending than men. In addition, the resulting comparative FPRs and FNRs indicate that the models provide both overestimates and underestimates of risk for women, which directly implies that they do not predict risk consistently for them. These conclusions reflect structural differences between the two groups' criminal histories, which indicate far fewer women reoffend, while the corresponding predictions differ. Therefore, it is consistent to conclude that for all datasets, the models created to predict recidivism appear to have bias against the Female group by violating the Equalized Odds fairness criterion.

Next, we study model predictions with bias mitigation. To accomplish that task, we perform rigorous statistical inference to study the distributions reported in Figures 2b,d, 3b,d, 4b,d, 5b,d and 6b,d by considering only FPR and FNR values (i.e., we do not consider the DFPR and DFNR values).

The methodology to create the models is presented in Sections 2.1–2.3. In a similar fashion to the previously reported simulations, Table 9 depicts the populations' normality check test and Table 10 the inference statistics test. Again, acceptance of the null hypothesis of normality in Table 9 implies the use of the *t*-test inference statistics test in Table 10, while rejection of the normality check in Table 9 implies the use of the Mann–Whitney U inference statistics test in Table 10.

**Table 9.** Acceptance/rejection of the null hypothesis obtained by the Shapiro–Wilk normality check test for models created with bias mitigation.

| Data Set/Output Attribute | FPR for Males/Females | FNR for Males/Females |
|---|---|---|
| Bulgarian/Recidivism risk assessment | Reject | Accept |
| Greek/Recidivism within three 3 years | Reject | Reject |
| Greek/Recidivism | Accept | Accept |
| Portuguese/Recidivism within three 3 years | Reject | Accept |
| Portuguese/Recidivism | Accept | Reject |

In view of the above tables, it can be easily concluded that the differences in FRPs and FNRs between the Male and Female groups are significantly reduced in all simulation cases. In addition, it is worth noting that some *p*-values in Table 10 indicate that the null hypothesis is accepted, meaning that the bias has been fully mitigated. On the other hand, several *p*-values indicate that this hypothesis is rejected, which means that the mitigation process was not fully accomplished. Based on this observation, we strongly emphasize

the following remark. The use of the optimization process does not guarantee a global minimum. Thus, we intended to minimize the FPR and FNR differences to mitigate the bias and not to eliminate it. In that direction, Table 10 indicates that in some cases, optimization managed to detect near-global minimum solutions, while in other cases it failed to do so.

**Table 10.** Inference statistics results based on *t*-test (acceptance case in Table 6) or Mann–Whitney U (rejection case in Table 6) for models created with bias mitigation.

| Data Set/Output Attribute | FPR for Males/Females | | FNR for Males/Females | |
|---|---|---|---|---|
| | Statistical Metric | *p*-Value | Statistical Metric | *p*-Value |
| Bulgarian/Recidivism risk assessment | 949.5 | $1.22 \times 10^{-13}$ | −8.91653 | $1.31 \times 10^{-15}$ |
| Greek/Recidivism within three 3 years | 4436 | 0.0001686 | 5678 | 0.0009784 |
| Greek/Recidivism | −15.6009 | $2.59 \times 10^{-36}$ | 26.93084 | $4.05 \times 10^{-68}$ |
| Portuguese/Recidivism within three 3 years | 2344 | $8.67 \times 10^{-11}$ | 5773.5 | 0.058916 |
| Portuguese/Recidivism | −2.83209 | 0.005102 | 5038.5 | 0.926023 |

In any case, the differences in FNRs and FPRs have been substantially reduced. Therefore, it is consistent to conclude that the constrained optimization process imposed a strong effect and finally obtained the best possible results.

Finally, we focus on the trade-off between fairness and predictive performance. Figure 7 illustrates the mean values of the accuracies obtained by the 1D-CNN for the three datasets for the recidivism prediction case, and Figure 8 illustrates the respective values for the within 3 years recidivism prediction case, considering the models without and with bias mitigation.

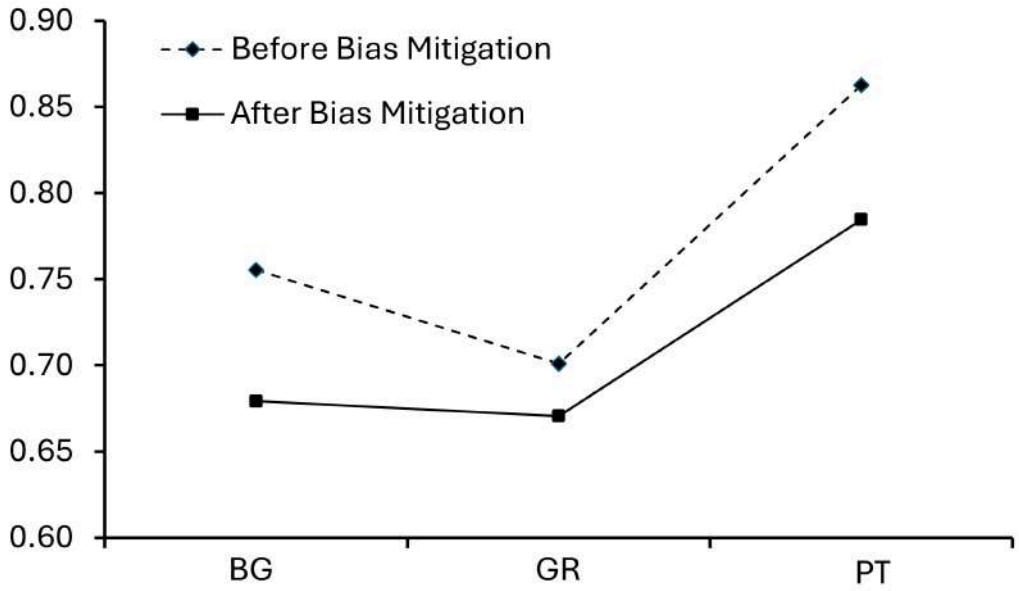


**Figure 7.** (Recidivism prediction case) Mean values of the accuracies obtained by the 1D-CNN for the three datasets (i.e., BG: Bulgarian data set, GR: Greek data set, PT: Portuguese data set) without and with bias mitigation.

Given the conclusions of the above analysis, it can be easily verified that the models' accuracies with bias mitigation are considerably smaller than the respective accuracies for the models without bias mitigation. Therefore, those figures directly quantify the trade-off between fairness and the accuracy performance of the models.

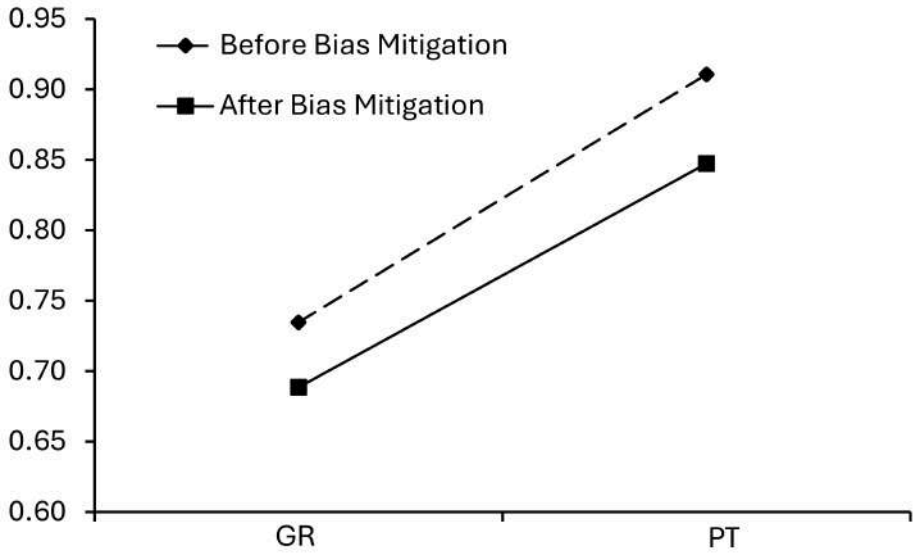


**Figure 8.** (Within 3 years recidivism prediction case) Mean values of the accuracies obtained by the 1D-CNN for the two datasets (i.e., GR: Greek data set, PT: Portuguese data set) before and after bias mitigation.

### *4.3. Experiment 3: Implementation and Study of the Predictive Parity Fairness Criterion*

In this section, we analyze the Predictive Parity (PP) fairness criterion and its impact on developing fair 1D-CNN classifiers for the three datasets. This criterion attempts to assess whether an ML model achieves equal positive predictions across the privileged and unprivileged groups. As such, the main requirement is to obtain equal values of the Positive Predicted Value (PPV) across groups. The current analysis acts as a supplement to the previous analysis, which was based on the EO fairness criterion, and intends to integrate our understanding regarding bias estimation and mitigation.

Given the nomenclature of Section 2.2, the PPV is

$$PPV = P\left(\hat{Y} = 1 | Y = 1\right) = \frac{TP}{TP + FP} \tag{33}$$

where TP is the True Positive and FP the False Positive. We can easily derive the values of TP and FP using the next soft differentiable approach [52]

$$TP = \sum_{i=1}^{n} y_i \hat{y}_i \tag{34}$$

and

$$FP = \sum_{i=1}^{n} (1 - y_i) \hat{y}_i \tag{35}$$

Thus, the PPV can be approximated as indicated next

$$PPV = \frac{\sum\limits_{i=1}^{n} y_i \hat{y}_i}{\sum\limits_{i=1}^{n} \hat{y}_i + \rho} \tag{36}$$

where $\rho$ is a small positive number. As a result, in our case, the PP fairness criterion is expressed in terms of the next equation

$$P\left(\hat{Y} = 1 | Y = 1, S = 1\right) = P\left(\hat{Y} = 1 | Y = 1, S = 0\right) \tag{37}$$

which can be rewritten as

$$\frac{TP|_{S=1}}{TP|_{S=1} + FP|_{S=1}} = \frac{TP|_{S=0}}{TP|_{S=0} + FP|_{S=0}} \tag{38}$$

Thus, the resulting constrained condition is

$$\psi_{PPV}(w) = \left| \frac{TP|_{S=1}}{TP|_{S=1} + FP|_{S=1}} - \frac{TP|_{S=0}}{TP|_{S=0} + FP|_{S=0}} \right| \tag{39}$$

Recalling that the main task is to minimize the $L_{CE}(w)$ in Equation (21), the constraint optimization problem becomes

$$\begin{aligned} &\text{minimize } L_{CE}(w) \\ &\text{subject to } \Psi_{PPV}(w) = (\psi_{PPV}(w))^2 - \delta \leq 0 \end{aligned} \tag{40}$$

Again, the above constrained problem is formulated by minimizing the subsequent regularization approach

$$L(w) = L_{CE}(w) + \alpha \Psi_{PPV}(w) \tag{41}$$

To resolve the above problem, we use again the 1D-CNN, where the structure and the parameter selection are the same as reported at the beginning of Section 4. The only difference is that, in this case, the values for the parameter $\alpha$ were found equal to 5, 3, and 5 for the Bulgarian, Greek, and Portuguese datasets. Also, here, 100 runs were executed for each experimental simulation. Figures 9–11 depict the results obtained for the three datasets regarding the cases without and with bias mitigation.

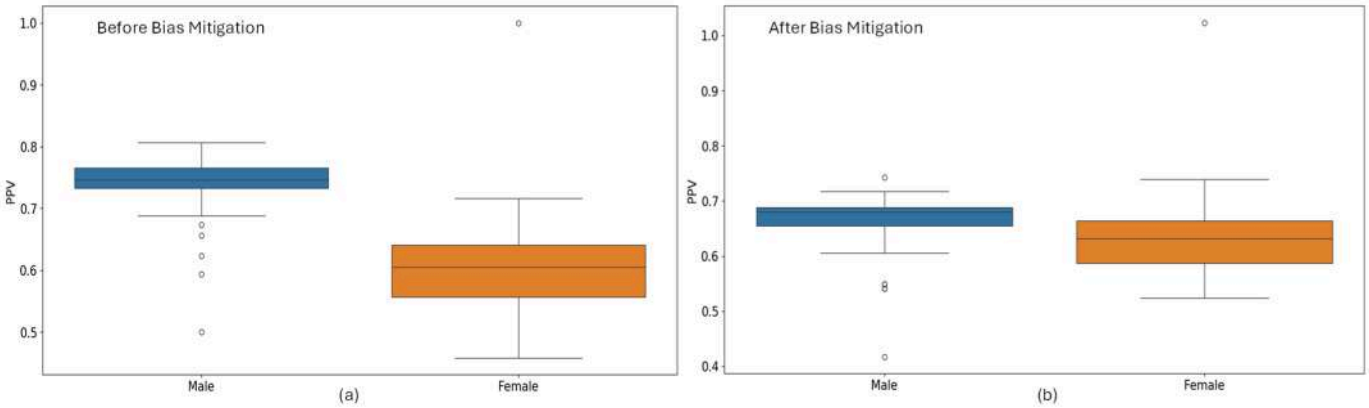


**Figure 9.** PPV values obtained by applying the Predictive Parity fairness criterion to the Bulgarian data set. Box-plots and the 95% confidence intervals for "Recidivism risk assessment" prediction case (100 runs for each model): (**a**) without bias mitigation, and (**b**) with bias mitigation.

In our recidivism case, the PP fairness criterion indicates the proportion of people who reoffended, out of all those the classifier predicts would do so. Thus, the detection of Predictive Parity imbalance is a strong indicator of the classifier's behavior related to groups with different characteristics [2]. In view of this remark, Figures 9a, 10a,c and 11a,c imply a direct imbalance regarding the PPV values for the Male and Female groups when no bias mitigation is taken into account. However, the imbalances reported in the above figures are not as strong as in Figures 2–6 for the "Before Bias Mitigation" cases.

On the other hand, Figures 9b, 10b,d and 11b,d directly indicate that the bias mitigation process managed to reduce the differences in PPV, yielding similar results as the ones reported in Figures 2–6 for the "After Bias Mitigation" case.

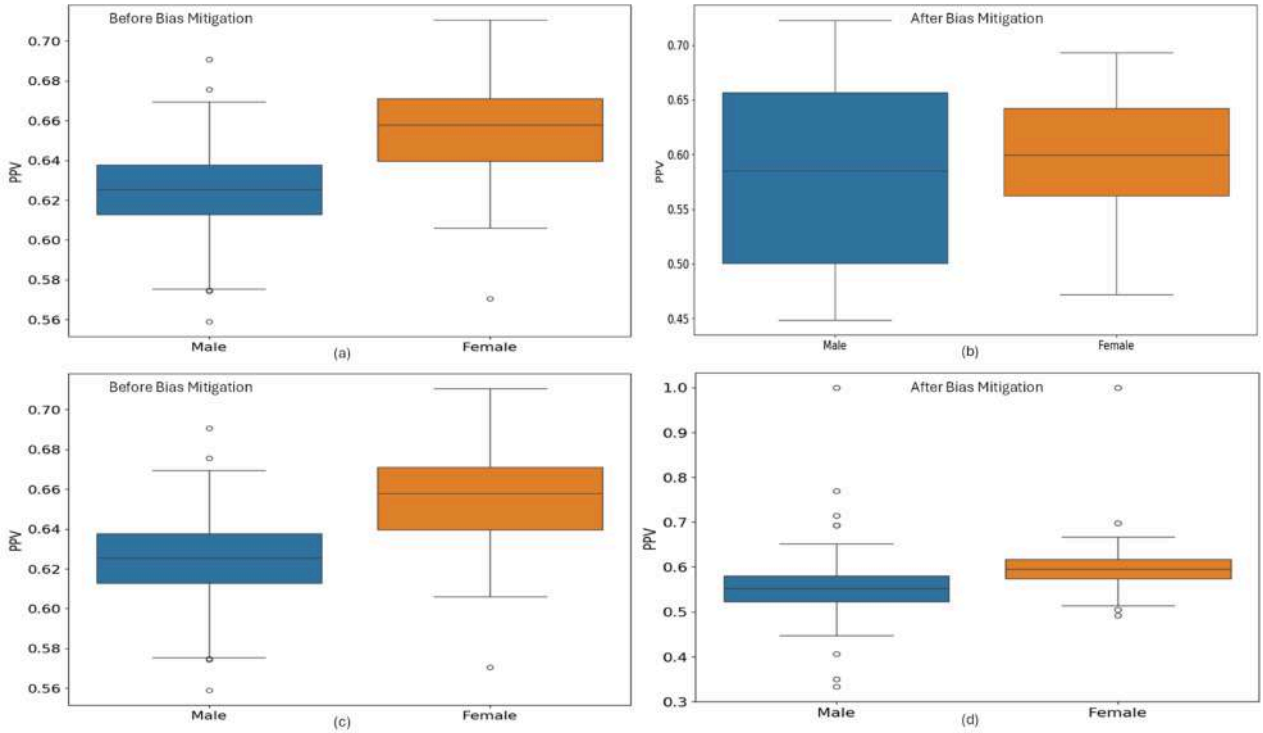


**Figure 10.** PPV values obtained by applying the Predictive Parity fairness criterion to the Greek data set. Box-plots and the 95% confidence intervals for "Recidivism" prediction case (100 runs for each model): (**a**) without bias mitigation, and (**b**) with bias mitigation. Box-plots and the 95% confidence intervals for "Recidivism within three 3 years" prediction case (100 runs for each model): (**c**) without bias mitigation, and (**d**) with bias mitigation.

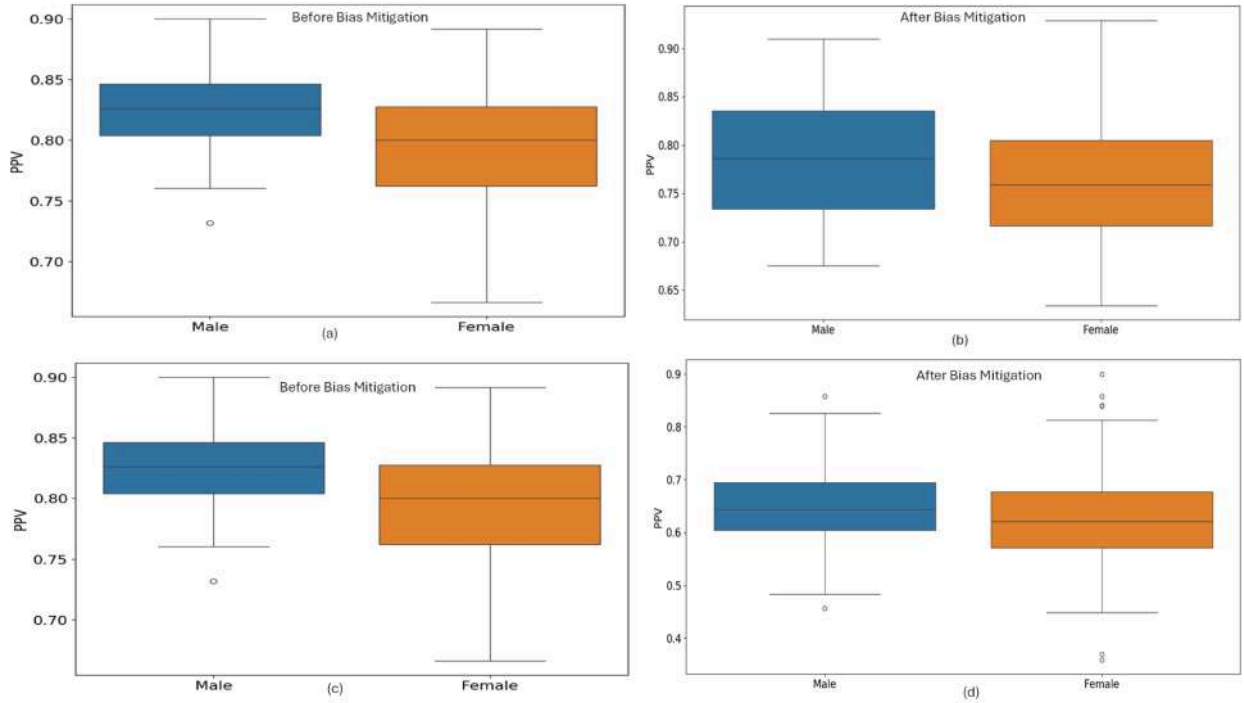


**Figure 11.** PPV values obtained by applying the Predictive Parity fairness criterion to the Portuguese data set. Box-plots and the 95% confidence intervals for "Recidivism" prediction case (100 runs for each model): (**a**) without bias mitigation, and (**b**) with bias mitigation. Box-plots and the 95% confidence intervals for "Recidivism within three 3 years" prediction case (100 runs for each model): (**c**) without bias mitigation, and (**d**) with bias mitigation.

Considering the implementation of EO and PP fairness criteria, we conclude that their combined effect strongly supports the assumption that the proposed methodology is in a position to produce fair recidivism classification predictions.

### *4.4. Kernel SHAP Configuration, Attribution Analysis, and Fairness Interpretation*

For the interpretability analysis, the Kernel SHAP explainer was configured as follows. A random background data set of 1000 instances, sampled without replacement from the training partition, was used to approximate the marginal feature distributions required for coalition value estimation, as described in Equation (28). This background size balances the trade-off between approximation fidelity and computational cost for the feature dimensionalities of the three datasets considered. For each explained instance, 2000 coalition samples were drawn to fit the weighted least-squares surrogate model in Equation (29); the sampling scheme follows the Kernel SHAP weighting function defined in Equation (30), which upweights singleton and full coalitions to improve the estimation of extreme marginal contributions. A logit link function was applied to the model's sigmoid output prior to attribution, so that the resulting SHAP values are additive in log-odds space, as expressed in Equation (32). Feature attributions were computed on the held-out test set for each of the 100 independent runs, and the resulting mean absolute SHAP values were aggregated across runs to obtain stable global importance rankings. All computations were performed using the KernelExplainer class, and check_additivity was enabled to verify local accuracy for each explained instance.

Figure 12 reports the grouped mean absolute SHAP values after aggregation of the one-hot encoded feature families. The dominant contributors are reducing punishment with work, sentence fulfilled, exemption from serving the sentence, and sentence for multiple crimes, followed by grouped marital status and grouped education level. By contrast, gender and nationality have negligible grouped importance, while days in prison and days of penalty have almost zero contribution at the global level. This result is important for the fairness analysis because it suggests that, in the fairness-constrained model, the global prediction structure is driven primarily by sentence-administration and legal-status variables rather than by direct use of the protected attribute.

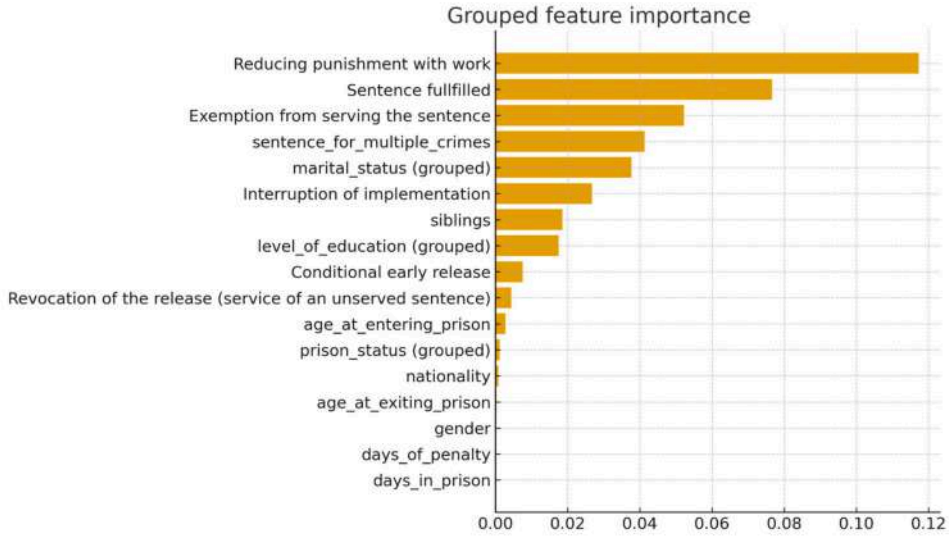


**Figure 12.** Grouped global SHAP importance for the fairness-constrained model, averaged across the 100 independent runs. One-hot encoded feature families are aggregated to provide a stable global ranking of feature influence.

Figure 13 provides a representative local waterfall explanation for a high-risk prediction. Starting from the baseline value E[f(X)] = 0.32, the predicted output increases to f(x) = 0.808 mainly due to the combined positive effects of non-exemption from serving the sentence, sentence fulfilled, reducing punishment with work, and sentence for multiple crimes. Conditional early release and the duration-related variables contribute only weakly, while gender does not appear among the dominant local drivers. Therefore, at the case level, the explanation is again dominated by operational and sentence-related attributes rather than by the protected characteristic itself.

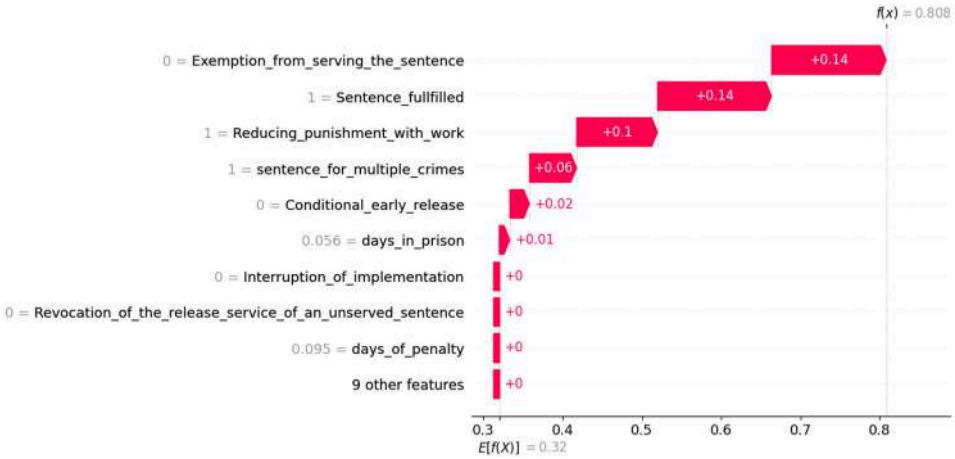


**Figure 13.** Representative local SHAP waterfall plot for a high-risk prediction produced by the fairness-constrained model. The plot decomposes the individual output into a baseline term and additive feature contributions in log-odds space.

The above explanations complement, rather than replace, the Equalized Odds analysis presented in Section 4.2. Equalized Odds quantifies fairness at the group level through disparities in false positive and false negative rates, whereas SHAP provides an audit of the feature pathways through which individual predictions are formed. In this sense, the weak direct contribution of gender in both the global and local explanations is consistent with the fairness-constrained training objective. The scientific value of the SHAP analysis lies in showing that the improved fairness outcomes are accompanied by a decision structure that places negligible direct weight on gender, while also revealing the operational variables through which residual disparities could still arise. This makes SHAP a complementary tool for fairness auditing, contestability, and model governance in criminal justice applications.

## 5. Conclusions

This study developed and evaluated a fairness-aware and interpretable framework for recidivism prediction, integrating a 1D Convolutional Neural Network, a custom Equalized Odds-constrained loss function, and Kernel SHAP-based explanations. The framework was applied to three distinct institutional datasets from Bulgaria, Greece, and Portugal, covering five prediction tasks and evaluated across independent runs per model, yielding statistically robust conclusions. The experimental results confirm that machine learning models trained without fairness constraints exhibit significant discriminatory behavior against the female group across all datasets and prediction horizons. Specifically, statistically significant differences in false positive rates and false negative rates between male and female offenders were detected in every baseline model, providing strong evidence that standard classification objectives reproduce and amplify structural biases present in historical criminal justice data. This finding holds regardless of national context or output definition, underscoring the systemic nature of the problem.

The fairness-constrained models achieved a substantial reduction in gender-based error rate disparities. In several experimental cases, the null hypothesis of equal distributions between Male and Female groups could not be rejected following bias mitigation, indicating full equalization of predictive treatment. Where statistically significant differences persisted, their magnitude was considerably diminished relative to the baseline. As expected, the introduction of fairness constraints involved a moderate reduction in overall classification accuracy, reflecting the genuine tension between unconstrained predictive optimization and equitable treatment across demographic groups. This trade-off is both theoretically anticipated and operationally manageable in the criminal justice context, where fairness and legitimacy are at least as important as aggregate performance.

Kernel SHAP analysis provided case-level and global interpretability of the constrained models, identifying the relative contribution of static attributes such as age, criminal history, and offense characteristics as the dominant drivers of individual risk predictions across all three national datasets. This finding is consistent with established criminological theory and provides a basis for auditing model behavior, supporting contestability, and informing institutional oversight. Crucially, interpretability was embedded as a structural principle of the framework rather than appended as a post hoc visualization layer, aligning with the requirements of trustworthy AI deployment in high-stakes settings.

Taken together, these results demonstrate that fairness, interpretability, and competitive predictive performance can be simultaneously pursued within a unified design framework for structured criminal justice data.

Future efforts will extend this framework to consider additional protected attributes such as nationality, employment/unemployment status, and age. Also, we will focus on examining cross-jurisdictional transferability more systematically, incorporating dynamic predictors as their longitudinal recording improves, and exploring the integration of counterfactual reasoning to further support procedural rights and practical contestability in deployment contexts.

**Supplementary Materials:** The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/a19070509/s1, Portuguese dataset used in the experiments in csv format and the variables' explanation in pdf format.

**Author Contributions:** Conceptualization, G.E.T. and S.C.; methodology, G.E.T. and S.C.; software, A.G.-R. and A.R.; validation, G.E.T., S.C., A.R., A.G.-R., E.V. and A.S.; formal analysis, G.E.T. and S.C.; investigation, G.E.T. and S.C.; resources, G.E.T. and S.C.; data curation, G.E.T., S.C., A.R. and A.G.-R.; writing—original draft preparation, G.E.T. and S.C.; writing—review and editing, E.V. and A.S.; visualization, K.K.; project administration, K.K.; funding acquisition, K.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was conducted in the context of FAIR-PReSONS project (https://fair-presons.aegean.gr/ accessed on 19 May 2026) Project ID: 101160473, funded by the European Union within e-JUSTICE program (https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/projects-details/43252386/101160473 accessed on 19 May 2026).

**Data Availability Statement:** The Portuguese dataset presented in the study is available in the Supplementary Materials. The Greek and Bulgarian datasets are available from the corresponding authors upon request.

**Acknowledgments:** We would like to thank the anonymous reviewers for their constructive comments and suggestions. We would like also to thank Algorithms' editorial team for the comments and suggestions. The authors reviewed all suggestions and took full responsibility for the final text.

**Conflicts of Interest:** Author Alvaro Garcia-Recuero was employed by the company "IPS Innovative Prison Systems & ICJT Innovative Criminal Justice Technologies". He participated in software, validation and data curation in the study. Author Eleni Valari and Andreas Siafakas were employed

by the company "IOTAM Ltd—Internet of Things Applications and Multi Layer Development". They participated in validation, writing—review and editing in the study. The roles of the two companies were participants in the project FAIR-PReSONS which was funded by the European Union. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| COMPAS | Correctional Offender Management Profiling for Alternative Sanctions |
| FPR | False Positive Rate |
| FNR | False Negative Rate |
| 1D-CNN | 1-Dimensional Convolutional Neural Network |
| CNN | Dimensional Convolutional Neural Network |
| Conv1D | 1-Dimensional Convolutional layer |
| Arnold PSA | Arnold Public Safety Assessment |
| SMOTE | Synthetic Minority Over-sampling Technique |
| AUC | Area Under the Curve |
| ReLU | Rectified Linear Unit |
| EO | Equalized Odds |
| ML | Machine Learning |
| TPR | True Positive Rate |
| SHAP | SHapley Additive exPlanations |
| DFPR | Difference in False Positive Rates |
| DFNR | Difference in False Negative Rates |

## References

1. Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* **2017**, *5*, 153–163. [CrossRef] [PubMed]
2. Oikonomou, F.; Bailis, E.; Bentos, B.; Chatzistamatis, S.; Tzortzi, M.; Kotis, K.; Spirou, S.; Tsekouras, G.E. Towards fair recidivism prediction: Addressing bias in machine learning for the Greek prison system. In Proceedings of the 5th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), Fez, Morocco, 15–16 May 2025.
3. Travaini, G.V.; Pacchioni, F.; Bellumore, S.; Bosia, M.; De Micco, F. Machine Learning and Criminal Justice: A Systematic Review of Advanced Methodology for Recidivism Risk Prediction. *Int. J. Environ. Res. Public Health* **2022**, *19*, 10594. [CrossRef] [PubMed]
4. Zeng, J.; Ustun, B.; Rudin, C. Interpretable classification models for recidivism prediction. *J. R. Stat. Soc. Ser. A* **2017**, *180*, 689–722.
5. Eaglin, J.M. Constructing recidivism risk. *Emory Law J.* **2017**, *67*, 59–122.
6. Feuerbach, L.; Skaramuca, D. The role of artificial intelligence in predicting recidivism. *Chall. Int. Crim. Crim. Law* **2025**, *2*, 271–294. [CrossRef]
7. Dressel, J.; Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.* **2018**, *4*, eaao5580. [CrossRef] [PubMed]
8. De la Cruz, R.; Padilla, O.; Valle, M.A.; Ruz, G.A. Modeling recidivism through Bayesian regression models and deep neural networks. *Mathematics* **2021**, *9*, 639. [CrossRef]
9. Farayola, M.M.; Bendechache, M.; Saber, T.; Connolly, R.; Tal, I. Enhancing algorithmic fairness: Integrative approaches and multi-objective optimization application in recidivism models. In *Proceedings of the ARES 2024*; ACM: New York, NY, USA, 2024.
10. Farayola, M.M.; Tal, I.; Connolly, R.; Saber, T.; Bendechache, M. Ethics and trustworthiness of AI for predicting the risk of recidivism: A systematic literature review. *Information* **2023**, *14*, 426. [CrossRef]
11. Farayola, M.M.; Tal, I.; Bendechache, M.; Saber, T.; Connolly, R. Fairness of AI in predicting the risk of recidivism: Review and phase mapping of AI fairness techniques. In *Proceedings of the ARES 2023*; ACM: New York, NY, USA, 2023.
12. Cavus, M.; Benli, M.N.; Altuntas, U.; Sari, M.; Ayan, H.; Ugurluoglu, Y.F. Transparent and bias-resilient AI framework for recidivism prediction using deep learning and clustering techniques in criminal justice. *Appl. Soft Comput.* **2025**, *176*, 113160. [CrossRef]
13. Tagliafierro, F.; Caterino, C. Criminal recidivism: Towards reliable and transparent predictive models. *Riv. Ital. Econ. Demogr. Stat.* **2026**, *80*, 367–378. [CrossRef]
14. Rebitschek, F.G.; Gigerenzer, G.; Wagner, G.G. People underestimate the errors made by algorithms for credit scoring and recidivism prediction but accept even fewer errors. *Sci. Rep.* **2021**, *11*, 20171. [CrossRef] [PubMed]

15. Cyphert, A.B. Reprogramming recidivism: The First Step Act and algorithmic prediction of risk. *Seton Hall Law Rev.* **2020**, *51*, 331–382.
16. Leng, J.; Xu, W.; Li, T.; Chen, L.; Xu, M. A prediction model of recidivism of specific populations based on Big Data. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 9167590. [CrossRef]
17. Ingram, E.; Gursoy, F.; Kakadiaris, I.A. Accuracy, fairness, and interpretability of machine learning criminal recidivism models. In *2022 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT)*; IEEE: Vancouver, WA, USA, 2022.
18. Ma, Y.; Nakamura, K.; Lee, E.-J.; Bhattacharyya, S.S. EADTC: An approach to interpretable and accurate crime prediction. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*; IEEE: Prague, Czech Republic, 2022.
19. Zhang, J. Research on the criminal recidivism prediction based on machine learning algorithm. In *Business, AI and Data Science 2022*; Atlantis Press: Dordrecht, The Netherlands, 2023; pp. 1297–1306.
20. Sultana, S.; Jahir, I.; Suukyi, M.; Nabil, M.M.R.; Waziha, A.; Momen, S. Advancing recidivism prediction for male juvenile offenders: A machine learning approach applied to prisoners in human province. In *Computational Methods in Systems and Software 2023*; Springer: Cham, Switzerland, 2024; Volume 935, pp. 184–201.
21. Scaria, A.G.; Subramanian, V.; George, N.K.; Sengupta, N. Algorithms and recidivism: A multi-disciplinary systematic review. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024), San Jose, CA, USA, 21–23 October 2024; pp. 1292–1305.
22. Asami, M. Rethinking Recidivism Risk Assessment Tool: From Prediction to Policy Learning, 2025. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5078134 (accessed on 19 May 2026).
23. Zhang, Y. Recidivism prediction model based on logistic regression. In *Proceedings of the CAMMIC 2025*; ACM: New York, NY, USA, 2025.
24. Guo, S. Recidivism prediction: A machine learning approach for a more efficient justice system. *Intell. Decis. Technol.* **2025**, *19*, 3304–3322. [CrossRef]
25. Lee, Y.; O, S. Redefining recidivism prediction: The impact of race and geographic location in machine learning models. *Crime Delinq.* **2025**, *71*, 3991–4017. [CrossRef]
26. Cohen, T.R.; Fronk, G.E.; Kiehl, K.A.; Curtin, J.J.; Koenigs, M. Clarifying the relationship between mental illness and recidivism using machine learning: A retrospective study. *PLoS ONE* **2024**, *19*, e0297448. [CrossRef] [PubMed]
27. Liu, J.; Li, D.M. Is machine learning really unsafe and irresponsible in social sciences? Paradoxes and reconsideration from recidivism prediction tasks. *Asian J. Criminol.* **2024**, *19*, 143–159. [CrossRef]
28. Jain, B.; Huber, M.; Fegaras, L.; Elmasri, R.A. Singular race models: Addressing bias and accuracy in predicting prisoner recidivism. In *Proceedings of the PETRA 2019*; ACM: New York, NY, USA, 2019.
29. Rudin, C.; Wang, C.; Coker, B. The age of secrecy and unfairness in recidivism prediction. *Harv. Data Sci. Rev.* **2020**, *2*. [CrossRef]
30. Wang, C.; Han, B.; Patel, B.; Rudin, C. In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *J. Quant. Criminol.* **2023**, *39*, 519–581.
31. Athota, J.K.; Parimi, K.K.; Teja, M.K.; Bhavani, M.A.; Devi, M.M.Y. Fairness in predicting recidivism score. In *Smart Data Intelligence*; Springer: Singapore, 2024; pp. 239–252.
32. Verrey, J.; Neyroud, P.; Sherman, L.; Ariel, B. A fairness scale for real-time recidivism forecasts using a national database of convicted offenders. *Neural Comput. Appl.* **2025**, *37*, 21607–21657. [CrossRef] [PubMed]
33. Farayola, M.M.; Tal, I.; Saber, T.; Connolly, R.; Bendechache, M. A fairness-focused approach to recidivism prediction: Implications for accuracy, trust, and equity. *AI Soc.* **2025**, *41*, 2783–2801. [CrossRef]
34. Sushiiel, A. Nam++: Achieving interpretability and fairness without sacrificing accuracy through neural additive models with selective feature interactions. *J. Electr. Syst. Inf. Technol.* **2026**, *13*, 21. [CrossRef]
35. Romero, M.A.; Orizaga Trejo, J.A.; Hernendez Mota, D.; Baltazar Villalpando, L.F.; Cruz Herrera, M.H. Enhancing explainability, privacy, and fairness in recidivism prediction through local LLMs and synthetic data. *Int. J. Comb. Optim. Probl. Inform.* **2025**, *16*, 60–70. [CrossRef]
36. Portela, M.; Castillo, C.; Tolan, S.; Karimi-Haghighi, M.; Pueyo, A.A. A comparative user study of human predictions in algorithm-supported recidivism risk assessment. *Artif. Intell. Law* **2025**, *33*, 471–517.
37. Chen, F.; Hou, H. Recidivism Prediction: A novel machine learning-based imbalanced learning method combined with the differential equation algorithm. *J. Appl. Sci. Eng.* **2026**, *29*, 765–780.
38. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef] [PubMed]
39. Yang, W.; Lorch, L.; Graule, M.A.; Lakkaraju, H.; Doshi-Velez, F. Incorporating interpretable output constraints in bayesian neural networks. In *Advances in Neural Information Processing Systems 33*; NeurIPS: Vancouver, BC, Canada, 2020.
40. Azeroual, A.; Taher, Y.; Nsiri, B. Recidivism forecasting: A study on process of feature selection. In *Proceedings of the NISS 2020*; ACM: New York, NY, USA, 2020.

41. Ning, Y.; Ong, M.E.H.; Chakraborty, B.; Goldstein, B.A.; Ting, D.S.W.; Vaughan, R.; Liu, N. Shapley variable importance cloud for interpretable machine learning. *Patterns* **2022**, *3*, 100452. [CrossRef] [PubMed]
42. Sun, J.; Shen, T. AI-assisted sentencing modeling under explainability constraints: Framework design and judicial applicability analysis. *Information* **2026**, *17*, 234. [CrossRef]
43. Sode, K. The interpreter's trap: How explainable AI launders uncertainty into justification-a socio-legal case study of COMPAS risk assessment. *AI Soc.* **2026**, *41*, 5285–5299. [CrossRef]
44. Skeem, J.; Lowenkamp, C. Using algorithms to address trade-offs inherent in predicting recidivism. *Behav. Sci. Law* **2020**, *38*, 259–278. [CrossRef] [PubMed]
45. Jin, Y.; Zheng, X.; Guo, L. Adaptive sentencing prediction with guaranteed accuracy and legal interpretability. *arXiv* **2025**, arXiv:2505.14011.
46. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.
47. Mu, D.; Zhang, S.; Zhu, T.; Zhou, Y.; Zhang, W. Prediction of recidivism and detection of risk factors under different time windows using machine learning techniques. *Soc. Sci. Comput. Rev.* **2024**, *42*, 1379–1402. [CrossRef]
48. Kiranyaz, S.; Avci, O.; Abdeljaber, O.; Ince, T.; Gabbouj, M.; Inman, D.J. 1D convolutional neural networks and applications: A survey. *Mech. Syst. Signal Process.* **2021**, *151*, 107398. [CrossRef]
49. Lee, Y.; O, S.; Eck, J.E. Improving recidivism forecasting with a relaxed naive Bayes classifier. *Crime Delinq.* **2025**, *71*, 89–117.
50. Circo, G.M.; Wheeler, A.P. An open source replication of a winning recidivism prediction model. *Int. J. Offender Ther. Comp. Criminol.* **2025**, *69*, 438–453. [PubMed]
51. Manisha, P.; Gujar, S. FNNC: Achieving fairness through neural networks. *arXiv* **2020**, arXiv:1811.00247v3.
52. Han, D.; Moniz, N.; Chawla, N.V. AnyLoss: Transforming classification metrics into loss functions. *arXiv* **2024**, arXiv:2405.14745.


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.