# Confronting Label Indeterminacy in Automated Bail Decisions

Cor Steging
c.c.steging@rug.nl
Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence, University of Groningen
The Netherlands

Tadeusz Zbiegień
tadeusz.zbiegien@doctoral.uj.edu.pl
Department of Legal Theory, Jagiellonian University
Poland

## Abstract

Bail decisions present a fundamental challenge for data-driven decision support systems. When bail is denied, the counterfactual outcome of whether the defendant would have appeared in court remains unobserved. As a result, historical bail data embed structural label indeterminacy: future decisions are influenced by past decisions whose outcomes are only partially knowable. Building automated systems on such data risks introducing bias and reinforcing feedback loops. This raises a core question for machine-learning systems intended to assist judicial actors: how should cases in which bail was denied be treated during model development? In a case study of bail decisions from the Unified Judicial System of Pennsylvania, we evaluate five contemporary approaches to handling label indeterminacy across three machine learning models, including a novel label imputation method motivated by the dynamics of bail decisions. Each method relies on unverifiable assumptions, yet all influence the models' predictive behaviour, sometimes even more so than the choice of model itself. Explainable AI analysis further reveals that these effects extend to the models' internal decision-making processes as well. Finally, we consider the notion of label indeterminacy from a legal perspective and assess the legitimacy of these approaches in the context of bail decision-making.

## Keywords

Label indeterminacy, Machine Learning, Bail decisions

## 1 Introduction

In the United States, more than two-thirds of the jail population consists of individuals who are legally presumed innocent and held in pretrial detention, a figure that continues to grow and imposes substantial social and financial costs on defendants and taxpayers alike [16]. Decisions about whether to grant bail or to detain a defendant pretrial therefore constitute a central challenge in the criminal justice system.

Technological decision support systems have been introduced to make bail decision-making more efficient and, potentially, more rational and thus equitable [1]. These systems typically aim to predict the likelihood that a defendant will fail to appear in court or whether they are likely to pose a risk to public safety, thereby informing judicial bail decisions. A well-known, yet controversial example of computer programs that assist in bail predictions is COMPAS, a rule-based system originally designed to predict recidivism [4]. More recently, it has been argued that modern AI solutions may contribute to enhancing judges' actual, rich, and perceived trustworthiness in the context of bail decision-making [15].

At the same time, automated bail decision systems face significant challenges: models can overestimate risk due to reliance on outdated data, embed implicit moral bias, and project a misleading appearance of scientific objectivity that may normalize preventive detention [11]. In practice, this can result in systematically inflated risk estimates, unexamined normative assumptions regarding how risk scores are used, and increased judicial acceptance of preventive detention.

Furthermore, comparing automated decisions with human judgment is inherently difficult due to the absence of counterfactual data [9]. As illustrated in Figure 1, when a defendant is denied bail and detained pretrial, they will in practice always appear in court, leaving it unknown whether they would have appeared or posed a risk had bail been granted. Consequently, automated predictions in these cases can only be compared to human decisions, rather than evaluated against a normative ground truth of actual risk or appearance. Moreover, when training an automated AI system on historical bail cases, the absence of counterfactual information renders the labels of pretrial detention cases indeterminate, as court appearance is observed only as a consequence of the detention decision rather than the defendant's true underlying propensity to appear (see Figure 1). While the problem of counterfactuals is not new and prior work has shown that using such data without accounting for indeterminate labels can perpetuate historical bias [4], the challenge of handling these indeterminate labels in a machine learning context has yet to be addressed. With automated decision support systems increasingly influencing bail outcomes, properly addressing this issue is critical.

Methods exist to impute indeterminate labels, but all depend on unverifiable assumptions [17]. Prior work on label indeterminacy in the legal domain has focused on appeals, showing that different approaches for handling indeterminate labels can produce substantial variation in model behaviour [19]. In this study, we examine label indeterminacy in the context of bail decisions. We apply five label imputation methods to machine learning models trained to predict failure to appear and evaluate these methods from a legal perspective.

## 2 Background

Bail is usually understood as a conditional release of the accused before trial, allowing them to remain free pending court proceedings, usually on the condition that they agree to appear in court and sometimes also on the condition that they provide financial security. It aims to balance the presumption of innocence, personal liberty, public safety, and the obligation to appear in court [1]. A growing

This manuscript has been accepted for presentation as a short paper at the 21st International Conference of AI & Law in Singapore, June 8 to 12 of 2026.

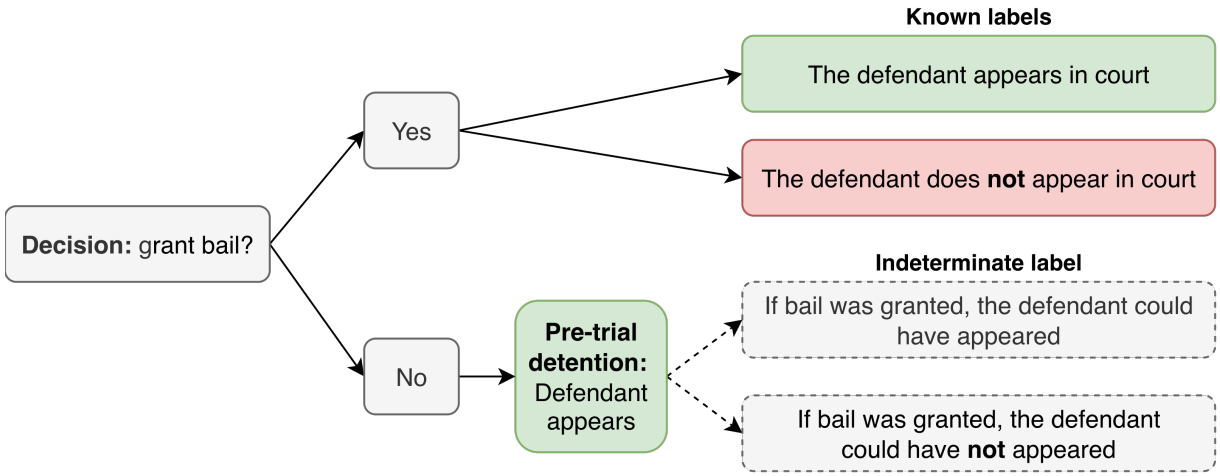



**Figure 1: Flowchart of the bail decision process illustrating how missing counterfactuals give rise to indeterminate labels.**

number of US states are eliminating cash bail, as it has been argued that it can disproportionately discriminate against defendants with limited financial resources, and are instead introducing risk assessment systems [1]. These systems rely on algorithmic models that take into account defendants' criminal history and other personal characteristics. However, the issue of pretrial assessment of defendants and their possible detention generates controversial problems that are faced not only by the United States but also by many legal systems around the world.

In the field of artificial intelligence and law, previous work has explored the use of artificial intelligence to support bail decisions [15]. [1] provides an analysis of pretrial risk assessment tools highlighting the arguments raised by both proponents and opponents. Research on New York City cases from 2008 to 2013 suggested potential improvements in bail decision-making [9], whereas a large randomized field trial found that providing judges with Public Safety Assessment (PSA) risk scores had little impact on detention outcomes [7].Related research has focused on predicting recidivism, which often informs pretrial risk assessment. The COMPAS tool has been used in the US for bail decisions [4], and the Department of Justice has reported on algorithmic approaches to recidivism prediction as well [5].

AI for automated bail decisions has also been explored in other jurisdictions. In India, fine-tuned language models [23] and large language models combined with retrieval-augmented generation (RAG) [18] have been applied to predict bail outcomes. In the UK, fairness-aware machine learning models have been shown to predict general and violent recidivism with high accuracy [21], while in Australia predictive modeling shows potential to support bail decisions but raises complex jurisprudential challenges [6].

Other work has examined the ethical and normative dimensions of AI in bail, including different conceptions of moral responsibility for AI and human agents [12]. Public attitudes toward AI-assisted judges have been shown to vary across racial groups, with judges relying on their own expertise generally rated more favorably than those relying on AI support [3].

The issue of counterfactuals is unavoidable in bail decision-making, since only one potential outcome is ever observed, with substantial implications for modeling and downstream applications [9]. Prior work has addressed this challenge by studying fairness in risk assessment tools under counterfactual settings and developing predictors that aim to ensure an equitable distribution of benefits and harms [14]. Other work introduces the concept of counterfactual loss to assess decisions based on all possible outcomes [10]. We, in turn, tackle the problem from the perspective of label indeterminancy.

Previous research on label indeterminacy has examined survival prediction in the medical domain, where decisions about withholding life-sustaining treatment are considered [17], and in the legal domain, focusing on overturned decisions in the European Court of Human Rights [19]. In both contexts, identifying indeterminate labels requires domain expertise, and methods have been explored and developed to impute these labels. Although all methods rely on unverifiable assumptions, differences in model behaviour are consistently observed depending on the imputation method employed.

## 3 Method

To investigate the effects of label indeterminacy in bail decisions, we train three machine learning models to predict defendant appearance while accounting for indeterminate cases where the outcome was determined by pretrial detention using five different label imputation methods. We apply each method and examine how it influences model behaviour. In this section, we describe the dataset, define label indeterminacy in the domain, outline the methods for handling indeterminate labels, and present the experimental setup.

### 3.1 Dataset

The dataset that we use consists of 90,732 publicly available and anonymized court cases from the Unified Judicial System of Pennsylvania, spanning January 2016 to June 2020 [22]. It is a tabular dataset that includes personal information about defendants, such as age and gender, as well as their criminal history, including the

number of felonies and misdemeanors. Case-specific information is also recorded, including the county of the court, the type of attorney, the type of bail requested (e.g., monetary), and the status of the bail request (e.g., denied).

The target feature in this dataset is whether a defendant appears in court or fails to appear (FTA). We predict this outcome by training machine learning models on the remaining features. In principle, these predictions could provide judges with additional information when deciding whether to grant bail.

Label indeterminacy is central to this task. When bail is denied, defendants almost always appear in court because they are held in pretrial detention. In our dataset, only 0.01% of bail-denied cases result in a defendant failing to appear, typically due to extreme events such as escape or death. This means that nearly all bail-denied cases produce a negative FTA label. However, this does not indicate whether the defendant would have appeared if bail had been granted, rendering these labels indeterminate. Machine learning models trained on such cases must therefore account for this indeterminacy.

Labels for denied cases are unambiguously indeterminate, but we classify several other case types as indeterminate as well. We define a label as indeterminate when an intervention alters the observed outcome such that it no longer reliably reflects the defendant's propensity to fail to appear. In all such cases, the defendant was detained pretrial, meaning the counterfactual outcome under release is unobserved.

On this basis, we also classify cases with bail status *Set* or *Partial Posting* as indeterminate. In these cases, defendants have not paid, or have only partially paid, their monetary bail, which typically results in pretrial detention and subsequently a court appearance in 99.3% of cases. The observed FTA outcomes thus reflect detention rather than the defendant's underlying willingness to appear. Similarly, the *Bond Terminated* status encompasses multiple permissible interventions (e.g., case resolution, dismissal, or transfer) that are not directly linked to appearance behaviour, yet typically result in court appearance due to pretrial detention (93.81%).

Conversely, we treat cases with the bail status *Posted*, indicating that bail has been paid or that alternative release conditions have been satisfied, as determinate. In these cases, defendants were released pretrial, and no intervention directly enforces court appearance or prevents failure to appear. Cases with the status *Forfeited* are likewise considered determinate. These cases involve defendants who were granted bail but failed to appear at a scheduled hearing, resulting in bail forfeiture and subsequent pretrial detention. The failure to appear is therefore directly observed. The same applies to cases with the status *Revoked* when revocation is explicitly due to a failure to appear. In contrast, revoked cases based on violations unrelated to failure to appear are treated as indeterminate, as bail loss in these instances is not causally attributable to non-appearance but instead arises from pretrial detention. Table 1 summarizes the determinate and indeterminate labels.

**Table 1: Label indeterminacy in the dataset.**

| | **Bail Status** |
|---|---|
| **Determinate** | Posted |
| | Forfeited |
| | Revoked (due to FTA) |
| **Indeterminate** | Denied |
| | Set |
| | Partial Posting |
| | Bond Terminated |
| | Revoked (other reasons) |

## 3.2 Preprocessing

To prepare the data for our experiments, we randomly sample 20% of the data, stratified by the FTA label, and set these aside as our test set. As the cases span 2016–2020, we assume temporal effects to be negligible. The test set consists of 18,147 cases, and the remaining 72,587 cases form the training set. In both datasets, approximately 3.8% of defendants failed to appear.

For training, we require a balanced label distribution, with 50% non-FTA and 50% FTA cases. As only 3.8% of the data involve FTA, we apply deterministic undersampling with full minority reuse to construct 25 balanced training sets. The FTA cases are identical across sets, while the non-FTA cases do not overlap. We will train a different model on each of these subsets and average the performance across the 25 subsets.

We remove six features from the data: the date, magistrate, bailType, bailAmount, bailStatus, and bailDenied. The first two are removed as they should be irrelevant to predicting the FTA label. The latter four are removed as they are information that is only available after bail has already been granted. As the predicted FTA likelihood of our models should be used to inform bail decisions, we cannot include these features. We decided to include all other features, including sensitive attributes such as race and sex, to faithfully reflect the information present in the data. Furthermore, exclusion would not remove their influence but instead shift it to correlated proxies, complicating both interpretation and bias assessment. We preprocess the data by midpoint-encoding binned numeric features such as age, ordinal-encoding ordered features like local employment levels, and one-hot encoding categorical variables, including race, sex, and type of attorney employed.

## 3.3 Imputing indeterminate labels

We examine five methods for imputing labels in indeterminate cases, as summarized in Table 2 and described in detail in this section. Each method is based on a defensible intuition but relies on unverifiable assumptions, so no single approach can be considered definitively correct. Four of the methods are based on previous work [17, 19], while the $daf$ approach is a novel method specifically motivated by the dynamics of bail decisions. We exclude approaches that require additional expert labeling, as used in other studies on label indeterminacy, to focus on methods that can be applied directly to existing datasets.

In the *Correct labels* ($corr$) approach, we use the training data as is with the original labels unchanged, implicitly assuming that defendants held in pretrial detention would have appeared in court had they been released.

In the *Detention-as-failure* ($daf$) approach, we also use the full training dataset but deterministically assign a failure-to-appear

**Table 2: Overview of the label imputation methods used to account for indeterminate labels in the bail domain.**

| ID | Name | Description | Assumption |
|---|---|---|---|
| *corr* | Correct labels | Includes indeterminate cases without altering their labels. | Assumes defendants who were in pretrial detention would have appeared in court if they were not detained. |
| *daf* | Detention-as-failure | Includes indeterminate cases and sets their label to 'fail to appear'. | Assumes defendants who were in pretrial detention would not have appeared in court if they were not detained. |
| *obs* | Observed only | Includes only cases where bail was not denied and the outcome is observable. | Assumes missing at random; fails if the choice for the denial of bail is based on the content of the case. |
| *obs + ip* | Observed + IP | Trains only on non-denied cases and applies inverse propensity weights to correct for sample bias | Assumes that every case has a non-zero chance of being granted bail (positivity) and that all factors influencing whether bail is granted are captured in the data (no unobserved confounders) |
| *nn* | Nearest neighbor | Imputes the labels of bail-denied cases using the most similar cases from the bail-granted sample | Assumes there is a valid similarity metric between bail-denied and bail-granted cases. |

label to all indeterminate cases involving pretrial detention. This assumes that the Magisterial District Judge's decision to order detention reflects an accurate expectation of non-appearance, and that defendants held pretrial would not have appeared in court had they been released.

The *Observed only* (*obs*) approach excludes all indeterminate cases, and trains models only on the determinate cases, as they have an observed label. This assumes that the indeterminate cases are a random sample of all cases.

The *Observed + IP* (*obs_ip*) approach corrects for sample bias from only observed labels by weighting each determinate case inversely to its probability of being observed, estimated via a propensity model (logistic regression) trained to distinguish determinate from indeterminate cases. This method assumes that the propensity model is correct and that all factors influencing bail decisions are captured in the dataset.

The *Nearest neighbor* (*nn*) method does not disregard the indeterminate cases, but instead imputes their FTA label based on similar determinate cases using a nearest neighbor algorithm with $k = 52$. This approach assumes that there is a valid similarity metric between cases, however.

As shown in Table 2, all methods rely on unverifiable assumptions, meaning that none can be considered definitively correct. In our experiments, we compare these methods and examine how they affect model behaviour.

### 3.4 Experimental setup

We evaluate the predictive behaviour of machine learning models trained to predict FTA based on case information. Three models are considered, all suitable for tabular bail decision data. First, we use logistic regression as a standard baseline. Second, following prior work on label indeterminacy in the medical domain, we include a random forest model [17]. Finally, we employ an XGBoost model, a widely used approach for tabular data that has also been applied to recidivism prediction [5]. Full model parameters are provided in the Appendix (Table 4).

Each model is trained on all 25 training subsets for each of the five label imputation methods. This results in a total of 348 trained models, for which we record predicted FTA probabilities on all cases in the test set.

## 4 Results

In Table 3, we present model performance on the test set, separated by determinate and indeterminate cases. We report the mean Matthew's Correlation Coefficient, scaled from -100 to 100, along with the standard deviation. Note that labels for indeterminate cases are treated as unknown in this study, so model performance on these cases should be interpreted cautiously. Likewise, performance on determinate cases reflects only a restricted subset (cases where bail was granted) and may not generalize to the full dataset.

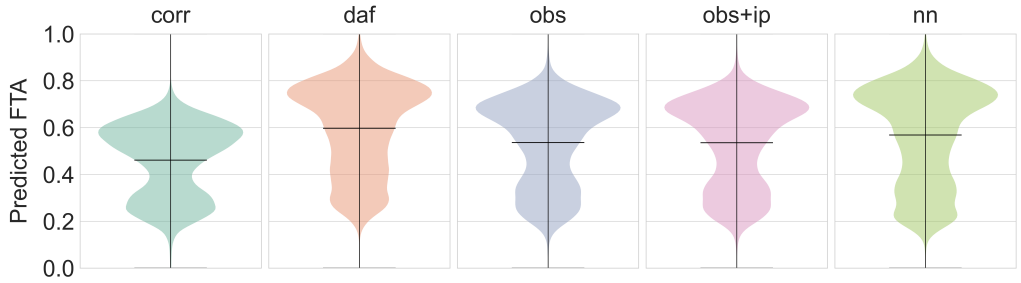
In Figure 2, we show the prediction distribution of each of the three models trained using each of the five label imputation methods, illustrating the predictive behaviour of the model.

We zoom in on the predictions for two individual cases in Figure 3, which shows the mean predicted FTA probabilities produced by the XGBoost model across the five label imputation methods.

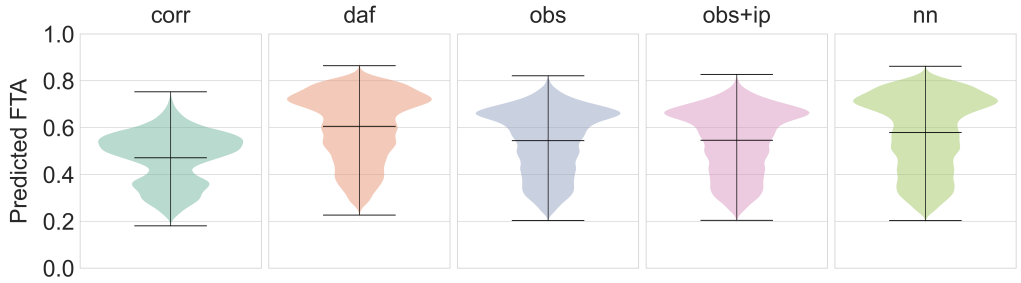
Additionally, we investigate how the five label imputation methods affect the internal decision-making of the models. In Figure 4, we report on the feature importance scores of the XGBoost classifier based on the gain metric. This metric reflects how much each feature contributes to improving the model's predictions when it is used to split the data during training. For each label imputation method, normalized gain scores are averaged across the 25 training subsets. We report the top 15 features ranked by normalized gain score in Figure 4.

**Table 3: Mean MCC and standard deviation for each model across the 25 training sets on determinate and indeterminate cases of the test set. These values should be interpreted with caution: they both reflect performance on only a restricted, unrepresentative subset of the data, and the true labels of the indeterminate cases are considered indeterminate.**

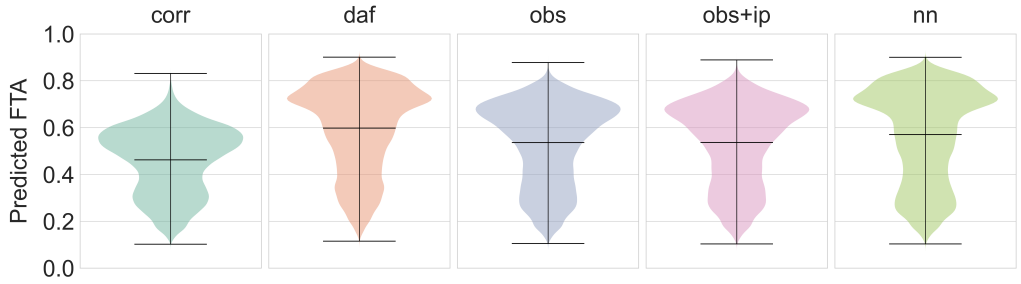
| Method | *corr* | *daf* | *obs* | *obs + ip* | *nn* |
|---|---|---|---|---|---|
| *Determinate cases* | | | | | |
| Logistic Regression | 10.78 ± 0.44 | 10.42 ± 0.23 | 10.85 ± 0.21 | 10.89 ± 0.22 | 10.70 ± 0.20 |
| Random Forest | 11.37 ± 0.31 | 9.79 ± 0.26 | 11.13 ± 0.17 | 11.06 ± 0.21 | 10.53 ± 0.27 |
| XGBoost | 10.61 ± 0.56 | 10.56 ± 0.30 | 11.41 ± 0.33 | 11.45 ± 0.31 | 10.99 ± 0.32 |
| *Indeterminate cases* | | | | | |
| Logistic Regression | 1.25 ± 0.58 | 1.00 ± 0.52 | 0.24 ± 0.50 | 0.60 ± 0.62 | 1.10 ± 0.59 |
| Random Forest | 2.51 ± 0.53 | 1.53 ± 0.16 | 1.73 ± 0.34 | 1.87 ± 0.43 | 1.74 ± 0.30 |
| XGBoost | 2.78 ± 0.52 | 1.52 ± 0.52 | 1.23 ± 0.61 | 1.15 ± 0.61 | 1.59 ± 0.73 |




**(a) Logistic Regression**




**(b) Random Forest**




**(c) XGBoost**

**Figure 2: Violin plots displaying the distribution of the predictions for the three models trained using different label imputation methods.**

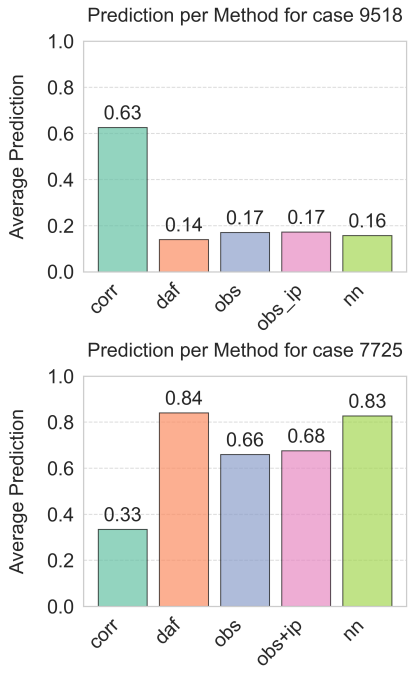



**Figure 3: The mean prediction of the XGBoost model for two specific cases using different label imputation methods.**

## 5 Discussion

We first present the experimental results, then examine the label imputation methods from a legal-philosophical perspective, and finally discuss the study's limitations and directions for future research.

### 5.1 Discussion of results

The violin plots in Figure 2 shows that the prediction distributions can vary quite significantly depending on the label imputation method used. The behaviour of the model is thus dependent on the way label indeterminacy is accounted for. This effect was also observed in the domain of overturned decisions in the European Court of Human Rights [19] and the medical domain [17]. The prediction distribution of the models trained using the *obs* and *obs_ip* method seem quite similar however, which was also seen in previous research. While the *obs_ip* aims to correct the sample bias of the *obs* method, seemingly no change can be observed in the resulting predictions of the models.

We confirmed these observations quantitatively using statistical tests (Wasserstein distance and Kolmogorov–Smirnov), which show that the differences between *obs* and $obs_ip$ are negligible, whereas *corr*, *daf*, and *nn* produce meaningfully distinct prediction distributions. While the violin plots in Figure 2 highlight differences in the overall shape and spread of predictions across models, the Wasserstein distances capture shifts in the distributions' absolute values across methods. Overall, both visual and quantitative analyses indicate that the choice of label imputation method substantially shapes model behaviour. Furthermore, further statistical analysis and a direct comparison of Wasserstein distances across all methods demonstrates that label imputation induces larger changes in predictions than the choice of model type.

Figure 3 shows the average predictions of the XGBoost model on two specific cases, further illustrating how the choice of label imputation method can shape the output of the model. To provide insight into the models' decision-making, Figure 4 presents the most important features identified by the XGBoost model. Certain features appear consistently across imputation methods; for instance, the leading offense and type of attorney rank among the top three features for all five methods, highlighting their strong predictive value. Overall, nine features appear in the top 15 across all methods, though their relative importances vary. An additional twelve features are present in the top 15 for only three, two, or a single method. These variations demonstrate that the five methods differ fundamentally, not just in their predictions, but also in how they utilize the case information in the decision-making process.

In Table 3, we report model performance separately for determinate and indeterminate test cases. These results should be interpreted with caution, since labels for indeterminate cases are inherently uncertain. For instance, a model predicting failure to appear for a case in which the judge ordered pretrial detention would, counterintuitively, be treated as an incorrect prediction. While the determinate cases have an observed label (the defendant was granted bail and either appeared or failed to appear), this subset of cases is subject to sample bias, as it only contains cases in which bail was granted. Therefore, the performance scores alone do not convey the insights we typically expect.

What we can see in this Table, however, is that the *obs*, $obs_ip$, and *nn* methods perform better on determinate cases than on indeterminate cases. In contrast, the *corr* method performs worst on determinate cases but best on indeterminate cases. This pattern arises because *corr* is the only approach that preserves the original labels of indeterminate cases. The other methods emphasize the observed determinate labels, explaining their stronger performance on determinate cases. This further illustrates how impactful the label imputation methods are on the predictive behaviour of the classifier.

### 5.2 Legal perspective on label imputation

Each of the five label imputation methods encodes a slightly different conception of justice in the context of risk governance. We argue that apart from purely technical consequences, design choices lead to normative commitments. The following subsection maps each method to its possible general legal-philosophical implications (See Table 2 for method definitions and assumptions; see Figure 1 for the role of counterfactuals in generating indeterminate labels). It should be emphasized that the following analysis does not aim to determine which of the examined imputation methods is normatively correct or preferable. Rather, we outline plausible lines of legal and philosophical arguments that may be advanced in relation to each approach. These arguments are not exhaustive, nor are they intended to be decisive.

The *corr* approach keeps the training data unchanged and includes denied-bail cases with their original recorded label, which is generally court appearances after detention. Defendants who

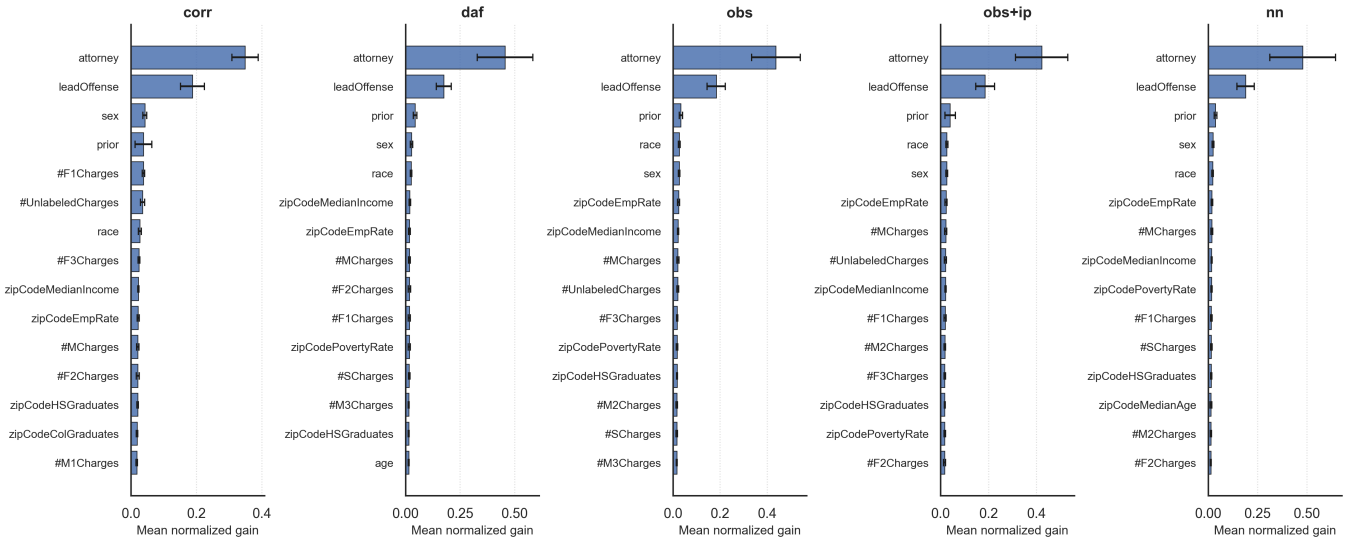



**Figure 4: Top 15 most important features of the XGBoost model per label imputation method, ranked by mean normalized gain, averaged across the 25 training subsets.**

were denied bail are assumed, for modeling purposes, to have appeared in court because they were detained, which gives rise to the 'appearance' label. In doing so, the *corr* method treats the resulting appearance labels as given and legitimate. Note that this might seem counter-intuitive, as the judge's decision to detain may intuitively equate to a 'failure to appear' label (which is what we describe in the *daf* method below). In the *corr* method, however, appearance is taken as evidence of reliability and the detention is treated as an adequate safeguard against non-compliance, rather than as an intervention that alters behaviour. The model effectively treats the compelled appearance as equivalent to voluntary compliance, which is a doubtful assumption to say the least. For machine-learning models, this implies that *corr* maps high-risk defendants, those who were denied bail and kept in pretrial detention, to 'likely to appear in court'. One could argue that normatively, this reflects a preventive-justice perspective and that it potentially even encodes that restrictive measures are accepted as substitutes for individualized risk assessment [1, 11, 13, 24]. If followed, this reasoning shifts responsibility away from the state by allowing liberty-restricting interventions to produce the very outcomes that justify them. Under *corr*, uncertainty about how defendants would have behaved if released is resolved by equating enforced compliance with genuine law-abiding conduct. As a result, the approach risks normalizing precautionary detention while hiding the underlying uncertainty.

The *daf* approach labels all indeterminate, pretrial-detention cases as FTA under the assumption that the prior judicial decision to detain correctly anticipated non-appearance. Whereas *corr* assumes that appearances due to pretrial detention reflect genuine compliance, *daf* treats past detention decisions as indicators of FTA. In legal-philosophical terms, under *daf* detention decisions are read as reliable indicators of the possible FTA, which, by consequence legitimizes precautionary detention as a proof of risk and empowers epistemic authority of past decision-makers [20]. It is easy to notice the risk that this could lead to a dynamic of self-fulfilling-prophecy loop [2]: detention-based predictions are trained on labels derived from earlier detention decisions, which are taken as a source of evidence regarding the risk of FTA, reinforcing and amplifying initial risk assessments over time in a loop. In recividity prediction, this has been shown to perpuate harmful bias [4]. Both the *corr* and *daf* approaches are questionable for prediction purposes, but they can potentially serve as lower and upper bounds, e.g., in the context of sensitivity testing.

The *obs* method limits modeling to cases where outcomes are observed, meaning cases in which bail was granted. Only appearances or failures to appear that occurred naturally, without pretrial detention, are treated as valid evidence. On the surface, *obs* avoids speculating about what might have happened in indeterminate, pretrial detention cases. In practice, however, it leaves past discretionary decisions largely unexamined by treating indeterminate cases as missing data. As a result, *obs* allows historical patterns of judicial discretion to persist. The model is trained on a selective sample and implicitly assumes that excluded cases are missing for neutral reasons [24]. This assumption is questionable, as in fact, judges' assessments of risk often shape which cases are granted bail. Under these conditions, *obs* can reproduce and reinforce existing biases, while appearing neutral by placing responsibility on the data rather than on the modeling choices.

Both *obs + ip* (inverse-propensity re-weighting) and *nn* (nearest-neighbor imputation) attempt to address missing outcomes through statistical adjustment or substitution. They rely on informed approximations rather than direct legal judgments. Method *obs+ip* re-weights observed cases to approximate the full population based on a specified propensity model, while *nn* assigns outcomes to detained defendants by comparing them to similar released defendants. Those approaches depend on strong and often untestable assumptions, such as whether the available variables are sufficient for estimating propensities or whether the chosen similarity measure is meaningful. These assumptions are rarely fully verifiable.

Consequently, one can also argue that the methods reflect a form of technocratic confidence in statistical tools to fill the gaps left by the legal process [13]. They assume that modeling can reliably replace missing observations. This could pose a risk in which weak probabilistic inference is turned into decisive evidence regarding individuals behaviour and, as a consequence, their freedom.

Furthermore, the statistical adjustment methods might pose a risk to individual contestability by affected individuals. An imputed FTA label is not a verifiable event. Defendants cannot readily challenge a counterfactual label, as they cannot demonstrate what would have occurred had bail been granted. Moreover, in practice there is a substantial risk that key properties of the model are not fully disclosed, such as how the data are preprocessed or which similarity metrics are used. At the same time, reliance on such models and different data imputation methods may redistribute responsibility across institutions, such as between judges and those developing or deploying a particular model [13, 24]. As a result, there is a risk that crucial factors in decision-making may no longer be rebuttable by those who are directly and most deeply affected by them.

With that being said, even when statistically defensible, different imputation methods can carry significant philosophical and legal implications. Accordingly, while we do not seek to resolve these debates or argue for or against the use of such models, we demonstrate that technical design choices operate within, rather than outside of, broader legal and moral discourses.

### 5.3 Limitations and future research

The fundamental challenge of label indeterminacy lies in determining which labels should be considered indeterminate. In our case, we focus on predicting failure to appear and therefore treat labels as indeterminate when the observed failure to appear outcome is affected by an intervention, most notably pretrial detention. Some additional questions arise regarding what should be considered indeterminacy. For example, when a judge sets monetary bail at a level that a defendant cannot afford, leading to pretrial detention, it is unclear whether this outcome should be treated as equivalent to a denial of bail and thus considered indeterminate, since one could argue that bail was technically granted. Defendants who are unable to pay monetary bail account for approximately one-third of the pretrial detention population in the United States [16], highlighting the practical importance of this issue. In our experiment, we treated cases in which bail was set but not paid as indeterminate, although the broader implications for how determinacy is defined warrant systematic investigation in future work.

Judges typically have access to bail reports containing substantially more information than is available in our dataset. Our models, by contrast, are limited to predicting failure-to-appear likelihoods and implicitly treat this as the sole basis for bail decisions, thereby excluding considerations such as public safety risk. This limitation does not affect the conceptual focus of the study, however, as the objective is not to develop a state-of-the-art classifier, but to illustrate how label indeterminacy arises in bail decision-making. For this reason, we focus on standard machine learning models that are relatively explainable and avoid more complex neural architectures. In practice, one could already argue that random forest and XGBoost models lack sufficient interpretability for such a highly sensitive task [8]. We therefore argue that symbolic or neuro-symbolic approaches may be more appropriate for this task, and that a human should always be kept in the loop. Additionally, we aim to examine how label imputation methods influence fairness metrics in future work. Finally, we did not include label imputation methods that require additional expert annotations and leave these for future research.

## 6 Conclusion

This paper examined the role of label indeterminacy in machine learning models for predicting failure to appear in pretrial bail decisions. A substantial portion of historical cases do not reliably encode defendants' appearance behaviour, as outcomes are often shaped by interventions such as pretrial detention. Treating these cases as fully informative introduces implicit and unverifiable assumptions that materially affect model behaviour.

Our empirical evaluation of multiple approaches to handling indeterminate labels across several machine learning models showed that methodological choices surrounding label treatment can have as much, or more, impact than model selection itself. Explainable AI analysis further revealed that these choices alter the internal decision-making of models, not just predictive performance. Moreover, each label imputation method is based on a set of unverifiable assumptions. We discuss how these assumptions carry subtle normative and legal implications and map each method to legal and philosophical arguments, highlighting how technical design choices operate within broader moral and legal discourses. These findings highlight label indeterminacy as a central concern with direct consequences for fairness, performance, and accountability in predictive models used for technological bail decision support systems.

## Acknowledgments

This research was funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl. The research leading to these results has received funding from the National Science Centre, Poland, project no. 2025/57/N/HS5/01561, titled "Uncertainty and Argumentation: Decision-Making Under Uncertainty in Legal Disputes."

## References

[1] D. Abu Elyounes. 2020. Bail or Jail? Judicial versus Algorithmic Decision-Making in the Pretrial System. *Science and Technology Law Review* 21, 2 (2020), 376–445.

[2] K. Bauer and A. Gill. 2024. Mirror, Mirror on the Wall: Algorithmic Assessments, Transparency, and Self-Fulfilling Prophecies. *Information Systems Research* 35, 1 (2024), 226–248.

[3] A. Fine, E. R. Berthelot, and S. Marsh. 2025. Public Perceptions of Judges' Use of AI Tools in Courtroom Decision-Making: An Examination of Legitimacy, Fairness, Trust, and Procedural Justice. *Behavioral Sciences* 15, 4 (2025), 476.

[4] A. W. Flores, K. Bechtel, and C. T. Lowenkamp. 2016. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation* 80 (2016), 38.

[5] T. Han. 2021. *Recidivism Forecasting Using XGBoost.* Technical Report NCJ 305033. U.S. Department of Justice, National Institute of Justice, Washington, D.C. https://nij.ojp.gov/library/publications/recidivism-forecasting-using-xgboost

[6] B. A. Hansard and J. Zhou. 2025. Jurisprudence and the Intelligible World: Exploring Predictive Modelling as a Mechanism to Decide Bail in the Australian Context. *International Annals of Criminology* 63, 3 (2025), 456–492.

[7] K. Imai, Z. Jiang, D. J. Greiner, R. Halen, and S. Shin. 2023. Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 186, 2 (2023), 167–189.

[8] T. Kavzoglu and A. Teke. 2022. Predictive Performances of Ensemble Machine Learning Algorithms in Landslide Susceptibility Mapping Using Random Forest, Extreme Gradient Boosting (XGBoost) and Natural Gradient Boosting (NGBoost). *Arabian Journal for Science and Engineering* 47, 6 (01 Jun 2022), 7367–7385.

[9] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. 2017. Human Decisions and Machine Decisions. *Q J Econ* 133, 1 (Aug. 2017), 237–293.

[10] B. Koch and Imaim K. 2025. Statistical Decision Theory with Counterfactual Loss. arXiv:2505.08908 [math.ST]

[11] J. L. Koepke and D. G. Robinson. 2018. Danger ahead: Risk assessment and the future of bail reform. *Washington Law Review* 93 (2018), 1725.

[12] G. Lima, N. Grgić-Hlača, and M. Cha. 2021. Human Perceptions on Moral Responsibility of AI: A Case Study in AI-Assisted Bail Decision-Making. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 235, 17 pages.

[13] C. McKay. 2020. Predicting Risk in Criminal Procedure: Actuarial Tools, Algorithms, AI and Judicial Decision-Making. *Current Issues in Criminal Justice* 32, 1 (2020), 22–39.

[14] A. Mishler, E. H. Kennedy, and A. Chouldechova. 2021. Fairness in Risk Assessment Instruments: Post-Processing to Achieve Counterfactual Equalized Odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 386–400.

[15] A. Morin-Martel. 2024. Machine learning in bail decisions and judges' trustworthiness. *AI & SOCIETY* 39, 4 (01 Aug 2024), 2033–2044.

[16] J.W.A. Ranson, A. N. Arnio, and J.E. Copp. 2023. Jurisdictional context and the (over)use of pretrial detention. *Social Science Research* 112 (2023), 102872.

[17] J. Schoeffer, M. De-Arteaga, and J. Elmer. 2025. Perils of Label Indeterminacy: A Case Study on Prediction of Neurological Recovery After Cardiac Arrest. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 1080–1094.

[18] P. K. Srivastava, U. Raj, P. Patel, S. K. Nigam, N. Shallum, and A. Bhattacharya. 2025. IBPS: Indian Bail Prediction System. *ArXiv* abs/2508.07592 (2025).

[19] C. Steging and T. Zbiegień. 2025. Label Indeterminacy in AI & Law. In *Legal Knowledge and Information Systems - JURIX 2025: The Thirty-eighth Annual Conference on Legal Knowledge and Information Systems (Frontiers in Artificial Intelligence and Applications, Vol. 416)*, R. Markovich, L. Di Caro, A. Rapp, and C. Schifanella (Eds.). IOS Press, Turin, Italy, 364–370.

[20] A. Stein. 2008. On the Epistemic Authority of Courts. *Episteme: A Journal of Social Epistemology* 5, 3 (2008), 402–410.

[21] J. Verrey, P. Neyroud, L. Sherman, and B. Ariel. 2025. A Fairness Scale for Real-Time Recidivism Forecasts Using a National Database of Convicted Offenders. *Neural Computing and Applications* 37, 26 (2025), 21607–21657.

[22] J. Williams and J. Z. Kolter. 2021. A Bayesian Model of Cash Bail Decisions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 827–837.

[23] E. Zacharia, W. Castelino, A. Puthran, and J. Mittal. 2025. Legal Insight - AI-Driven Bail Prediction System. *International Journal For Multidisciplinary Research* 7 (08 2025).

[24] A. Završnik. 2020. Criminal Justice, Artificial Intelligence Systems, and Human Rights. *ERA Forum* 20 (2020), 567–583.

# Appendix

## A Model hyperparameters

The full parameters of each model used in this study can be seen in Table 4. We include these for transparency and reproducibility reasons. Most settings were kept at default, or were inspired by previous research [5, 17]. The full code will be made available upon acceptance.

**Table 4: Hyperparameters used for each model**

| | |
|---|---|
| **Logistic Regression** | |
| Maximum iterations | 2000 |
| Regularization | $\ell_2$ |
| **Random Forest** | |
| Number of trees | 500 |
| Maximum tree depth | 8 |
| Minimum samples per leaf | 50 |
| Feature subsampling | $\sqrt{p}$ |
| **XGBoost** | |
| Objective | binary:logistic |
| Number of trees | 800 |
| Maximum tree depth | 4 |
| Learning rate ($\eta$) | 0.005 |
| Row subsampling | 0.8 |
| Column subsampling | 0.9 |
| $\ell_2$ regularization ($\lambda$) | 5.0 |
| $\ell_1$ regularization ($\alpha$) | 1.0 |
| Minimum child weight | 20 |