**Frontiers in Computer Science and Artificial Intelligence**
DOI: 10.32996/fcsai
Journal Homepage: www.al-kindipublisher.com/index.php/fcsai

FCSAI
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

**| RESEARCH ARTICLE**

# Artificial Intelligence in Sentencing: Evaluating Machine Learning Models for Sentencing Recommendations in the U.S.

**Mohammed Nazmul Islam Miah[1], Md Joshim Uddin[2] and Manohar Kakumani[3]**

[1]Master of Public Administration, Gannon University, Erie, PA, USA
[2]Master of Law, ASA University of Bangladesh.
[3]Master of Science in Computer and Information Science, Gannon University, Erie, PA, USA
**Corresponding Author**: Mohammed Nazmul Islam Miah, **Email**: islammia001@gannon.edu

**| ABSTRACT**

Artificial intelligence is increasingly deployed in high-stakes decision-making, raising critical questions about accuracy, fairness, and transparency in regulated domains. This study evaluates the use of machine learning models to generate sentencing recommendations within the U.S. criminal justice system, examining whether such models can reliably support judicial decision-making without amplifying existing inequities. Using a comprehensive dataset of sentencing records enriched with engineered features reflecting criminal history, offense severity, demographics, and jurisdictional context, we develop and compare a range of predictive models, including Logistic Regression, tree-based ensembles (Random Forest, XGBoost, LightGBM), deep learning architectures (MLP, LSTM, Bi-LSTM), and hybrid ensemble frameworks. Models are assessed on both continuous sentence length prediction and classification of above-median sentencing, using metrics such as mean absolute error, R-squared, AUC-ROC, and F1-score. Fairness metrics are computed across gender, age, and jurisdictional groups, and interpretability analyses employ feature importance, attention weights, and SHAP values to ensure transparency in decision logic. Results indicate that hybrid and stacked ensembles achieve the best balance of accuracy and fairness improvements over baselines, with interpretability tools confirming alignment with legal reasoning and risk factors. These findings suggest that responsibly governed AI systems can augment sentencing decisions as decision-support tools, provided continuous bias monitoring and ethical oversight are integrated into deployment practices. The study contributes empirical evidence and methodological guidance for integrating machine learning into judicial contexts.

**| KEYWORDS**

Artificial intelligence; machine learning; sentencing recommendations; criminal justice; fairness; interpretability.

**| ARTICLE INFORMATION**

**ACCEPTED:** 01 February 2026 **PUBLISHED:** 22 February 2026 **DOI:** 10.32996/fcsai.2025.4.3.5x

## 1. Introduction

### 1.1 Background and Motivation

Artificial intelligence (AI) has increasingly permeated high-stakes and regulated domains, demonstrating a remarkable capacity to process complex data streams and support decision-making in environments where errors carry significant consequences. In national-scale cybersecurity systems, Das et al. (2025) highlight how AI-driven threat detection frameworks enable real-time monitoring and automated response mechanisms, effectively enhancing resilience against sophisticated attacks [8]. Similarly, Debnath et al. (2025) show that energy systems benefit from AI-based anomaly detection, providing predictive safeguards in safety-critical infrastructure and illustrating AI adoption in environments where human oversight alone may be insufficient [9].

The widespread integration of AI across such high-impact sectors underscores the relevance of its governance, particularly in contexts where operational failures may result in systemic harm or loss of life.

Beyond infrastructure, AI has found applications in national security and education regulation, domains characterized by high societal stakes and intricate regulatory oversight. Miah et al. (Threat Intelligence, 2025) illustrate the deployment of machine learning algorithms for threat intelligence, demonstrating AI's utility in predicting and mitigating risks to national safety [17]. In parallel, Miah et al. (Education Regulation, 2025) emphasize the importance of algorithmic accountability in legal and educational oversight, arguing that AI systems, if unchecked, may perpetuate inequities or unintended harm [18]. Collectively, these studies provide a compelling rationale for scrutiny of AI interventions in any regulated, high-impact setting, including the judicial system, where decisions directly affect individual liberty and societal trust.

In criminal justice, AI adoption is particularly contentious. Tools such as COMPAS have demonstrated the potential for data-driven risk assessments, yet they also raise critical questions regarding fairness, transparency, and interpretability. Algorithmic recommendations for sentencing and parole decisions present an inherently high-stakes application; predictive errors or biases can result in disproportionate penalties or systematic discrimination. The combination of technical capability and ethical complexity positions AI in sentencing as an exemplar of the broader challenges observed in other critical infrastructure domains. By situating judicial AI within the larger ecosystem of regulated, high-impact applications, it becomes evident that careful evaluation of model performance, bias mitigation, and explainability is not only a technical imperative but also a societal necessity. This context underscores the study's focus on evaluating machine learning models for sentencing recommendations, aiming to inform both scholarly understanding and policy guidance in deploying AI responsibly within criminal justice systems.

### 1.2 Research Problem and Objectives

Predictive modeling in sentencing draws conceptual parallels to risk scoring frameworks in finance and macroeconomic early warning systems. Chouksey et al. (2025) provide a foundation in AI-driven early warning systems for financial risk, demonstrating how predictive analytics can anticipate critical events in complex, dynamic systems [6]. Similarly, Rahman (2025) illustrates the use of machine learning for detecting micro-inflation clusters across the U.S. economy, employing structured modeling approaches to identify emergent anomalies before they materialize into systemic disruptions [20]. Reza et al. (2025) further underscore the utility of real-time early warning models in financial distress prediction, emphasizing adaptive learning techniques that process heterogeneous data streams to generate actionable insights [23]. These frameworks collectively highlight the potential of predictive analytics to assess risk in forward-looking, high-stakes environments, offering methodological justification for their translation to the criminal justice context.

Sentencing recommendations represent a form of risk estimation where models attempt to predict recidivism probability, likely behavioral trajectories, or optimal intervention levels. Analogous to financial systems, the judicial domain must manage uncertainty, balance competing outcomes, and operate under regulatory and ethical constraints. Ray (2025) demonstrates the efficacy of multi-market crisis prediction through machine learning, employing diverse data sources to strengthen predictive robustness [21]. Translating this approach to sentencing involves analogous considerations: ensuring models are trained on representative datasets, assessing the influence of features such as criminal history, offense severity, and demographic factors, and validating outputs against both predictive accuracy and fairness criteria.

The central research question addressed by this study is whether machine learning models can provide reliable, unbiased, and interpretable sentencing recommendations. Specific objectives include: comparing predictive performance across alternative model architectures, evaluating fairness and bias across demographic groups, and assessing interpretability mechanisms to ensure that recommendations are transparent and actionable within legal and policy frameworks. The study's design draws from predictive modeling best practices established in risk-sensitive sectors, applying them to judicial decision support to illuminate potential trade-offs between accuracy, fairness, and interpretability. By establishing this methodological foundation, the research aims to advance both empirical understanding and practical guidance for responsible AI deployment in criminal justice, highlighting parallels with other high-stakes domains while addressing unique challenges inherent to legal decision-making.

### 1.3 Contributions of the Study

This study contributes to the growing literature on AI-assisted judicial decision-making by providing a systematic evaluation of machine learning models for sentencing recommendations. First, it conducts an empirical comparison of multiple predictive architectures, analyzing their performance in estimating recidivism risk and recommending appropriate sentencing ranges. Second, the study integrates fairness and explainability assessments, examining model outputs across demographic subgroups and applying interpretability techniques to ensure transparency in decision-support systems. Third, the work frames these contributions within governance and policy considerations, emphasizing the importance of responsible deployment, human oversight, and ethical accountability. Unlike prior work that often focuses solely on predictive accuracy or technical feasibility,

this research situates AI within a holistic framework where performance, fairness, and interpretability converge, providing actionable insights for policymakers, legal practitioners, and AI developers seeking to implement trustworthy, socially responsible judicial technologies.

## 2. Literature Review

### 2.1 AI in Criminal Justice and Risk Assessment

The application of artificial intelligence in criminal justice has attracted significant attention, particularly through the development of risk assessment instruments designed to predict recidivism and inform sentencing decisions. Angwin et al. (2016) conducted a landmark investigation into the widely used COMPAS system, revealing evidence of systematic racial disparities in predicted risk scores and sparking broad debate over the fairness and accountability of algorithmic sentencing tools [2]. Building upon these empirical critiques, Chouldechova (2017) provided a formal statistical analysis of disparate impact in recidivism prediction instruments, demonstrating that even models with ostensibly high predictive accuracy can produce unequal outcomes across demographic groups [5]. These studies collectively underscore the tension between predictive utility and fairness, highlighting the need for rigorous evaluation frameworks that account for both dimensions.

Dressel and Farid (2018) extended this line of inquiry by benchmarking algorithmic risk predictions against human judges' assessments, showing that machine learning models can achieve comparable or superior accuracy in forecasting recidivism while maintaining consistent decision criteria [10]. However, this work also revealed that predictive parity does not inherently ensure equity, reinforcing the importance of integrating fairness considerations into model evaluation. Rudin (2019) further argued for the adoption of interpretable models in high-stakes decision-making contexts, including criminal justice, emphasizing that transparency and explainability are not merely ethical luxuries but practical necessities to ensure accountability and public trust [24]. Collectively, these contributions establish a foundation for the development, evaluation, and critique of AI-based sentencing systems, highlighting the dual imperatives of accuracy and fairness.

The evolution from actuarial risk models to more sophisticated machine learning architectures has expanded the predictive capabilities of sentencing tools while simultaneously amplifying concerns over bias and opacity. Early actuarial approaches relied on linear scoring rules derived from historical recidivism data, often with limited feature representation and minimal consideration of interaction effects. Machine learning models, by contrast, can incorporate a broader array of inputs—including offense characteristics, prior criminal history, and contextual demographic information—enabling more nuanced risk stratification. Yet, as empirical studies demonstrate, these gains in predictive performance do not automatically translate into socially equitable outcomes. The literature suggests that effective deployment of AI in criminal justice requires a holistic perspective that integrates predictive efficacy with careful fairness assessment and interpretability mechanisms, providing a benchmark against which new models can be evaluated.

### 2.2 Algorithmic Fairness Theory and Bias

The theoretical underpinnings of fairness in machine learning provide critical guidance for assessing sentencing models. Barocas and Selbst (2016) laid the groundwork for understanding disparate impact in algorithmic systems, formalizing how data-driven models can reproduce or amplify social inequities even in the absence of explicit bias [3]. Binns (2018) complemented this work by exploring the philosophical foundations of fairness, emphasizing that fairness is inherently multidimensional and context-dependent, requiring careful consideration of societal norms and legal principles when designing algorithmic interventions [4]. Corbett-Davies et al. (2017) quantified the trade-offs associated with fairness constraints, demonstrating that optimizing for equity across groups often incurs measurable costs in predictive performance, a critical consideration for high-stakes judicial applications [7].

Hardt et al. (2016) introduced the concept of equality of opportunity in supervised learning, framing fairness as a constraint that ensures similar error rates across protected groups conditional on relevant outcomes [11]. Kleinberg et al. (2017) formalized inherent trade-offs in risk scoring, showing that certain fairness criteria, such as calibration and balance across groups, may be mutually incompatible in practical settings [14]. Kusner et al. (2017) extended these ideas with counterfactual fairness, proposing methods to evaluate whether algorithmic decisions would differ under hypothetical changes in protected attributes, thereby capturing causal notions of bias [15]. Finally, Selbst et al. (2019) emphasized the sociotechnical nature of fairness, arguing that ethical evaluation must extend beyond algorithmic outputs to include institutional context, regulatory environment, and the broader social ecosystem in which models operate [25].

These theoretical frameworks collectively inform the evaluation of sentencing algorithms by highlighting the complexity of bias mitigation and fairness assessment. In judicial AI systems, fairness cannot be treated as a single, static metric; rather, it requires ongoing measurement, multi-dimensional assessment, and contextualized interpretation. The literature underscores that predictive performance alone is insufficient; algorithmic interventions in high-stakes domains must integrate fairness principles

from both technical and sociological perspectives. By situating sentencing models within these fairness frameworks, researchers can more rigorously identify sources of bias, quantify their impact, and explore strategies for mitigation, ensuring that AI-assisted decision-making aligns with legal, ethical, and societal expectations.

### 2.3 Explainability and Legal Accountability

Interpretability and transparency are central to the responsible deployment of AI in judicial contexts. Lundberg and Lee (2017) introduced SHAP, a unified framework for attributing model predictions to individual features, providing a systematic approach to understanding complex models and enhancing trust in automated decisions [16]. Mitchell et al. (2019) proposed the concept of model cards, structured documentation that communicates model capabilities, limitations, and potential biases to stakeholders, thereby supporting accountability and informed oversight [19]. These interpretability tools are particularly relevant in high-stakes domains, where opaque recommendations could undermine due process or exacerbate public mistrust.

Sizan et al. (2025) demonstrated the importance of transparency in complex machine learning systems through their work on unsupervised anomaly detection in financial networks, illustrating how interpretability can reveal hidden patterns, validate model outputs, and provide evidence for auditing automated decisions [28]. In judicial AI, similar approaches are critical: interpretable models allow legal practitioners, policymakers, and oversight bodies to assess the rationale behind sentencing recommendations, verify compliance with ethical standards, and ensure that algorithmic decisions do not introduce unintended harm. Together, these contributions underscore that explainability is not merely an optional enhancement but a necessary component of AI systems deployed in socially sensitive, high-stakes environments.

## 3. Methodology

### 3.1 Data Collection and Preprocessing

The dataset for this study was constructed from publicly available U.S. sentencing records, spanning both state and federal court systems, to ensure a representative sample across jurisdictional contexts. These records include detailed information about defendants, offenses, and sentencing outcomes, providing a rich basis for machine learning analysis. Data collection focused on structured fields that could be reliably extracted and standardized, including criminal history, offense severity, demographic characteristics, and prior judicial interactions. To maintain data quality and analytical rigor, extensive preprocessing was performed. Missing or inconsistent values were identified and handled through a combination of imputation techniques and manual validation to ensure completeness without introducing artificial bias.

Feature engineering played a critical role in transforming raw court records into predictive inputs suitable for machine learning models. Key features included aggregated measures of criminal history, such as prior convictions and the severity of past offenses, as well as offense-specific attributes, including classification, statutory penalties, and contextual aggravating or mitigating circumstances. Demographic information, including age, gender, and jurisdictional identifiers, was encoded in a manner compatible with both categorical and numerical modeling frameworks. Interaction terms between criminal history and offense severity were introduced to capture nuanced risk patterns that might influence sentencing decisions. Temporal features, such as the elapsed time since prior convictions, were also derived to enhance the predictive capacity of models that account for recency effects in judicial decision-making.

To mitigate issues of class imbalance, which are prevalent in sentencing datasets due to the unequal distribution of offense types and sentencing outcomes, data balancing techniques were applied. Oversampling strategies were used for underrepresented classes, while outlier detection was employed to identify cases that could disproportionately influence model training. Standardization and normalization procedures were applied to numerical features to ensure that differences in scale did not bias learning algorithms. Categorical variables were encoded using one-hot and ordinal schemes as appropriate, preserving interpretability for downstream analysis. The final dataset thus combined rigorously cleaned, well-engineered, and balanced features, establishing a robust foundation for subsequent model development and evaluation.

#### Exploratory Data Analysis

Exploratory data analysis (EDA) was conducted to examine the distributional properties of key predictors and sentencing outcomes, as well as their interrelationships. Each visualization provided structured insight into feature behavior and informed subsequent model development decisions. The histogram of prior convictions reveals a strongly right-skewed distribution consistent with a Poisson process centered around one prior offense. The majority of defendants have zero, one, or two prior convictions, with frequencies declining sharply as the number of prior convictions increases. Only a small fraction of individuals exhibit more than four prior convictions, forming a long but thin tail. This distribution confirms that most defendants in the dataset have limited criminal histories, while a relatively small subset represents repeat offenders. The concentration at lower counts suggests that criminal history is highly imbalanced and may exert nonlinear effects in predictive modeling.

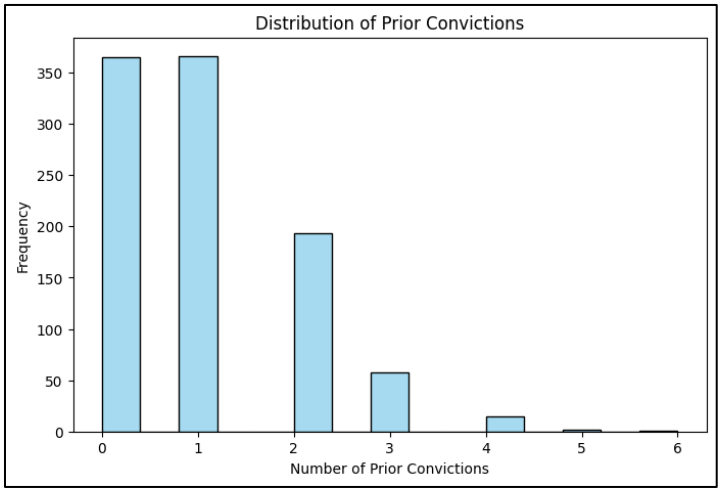


Fig.1: Distribution of Prior Convictions

The age histogram, with kernel density overlay, displays an approximately bell-shaped distribution centered in the mid-30s. The highest concentration of defendants falls between 25 and 40 years of age, with frequencies gradually tapering for both younger (18–24) and older (50+) individuals. The clipping at 18 and 70 years is evident from the bounded distribution. The smooth density curve confirms that age is approximately normally distributed within the dataset. This indicates that sentencing analyses are primarily influenced by individuals in early to mid-adulthood, while extreme age groups contribute less to overall variance.

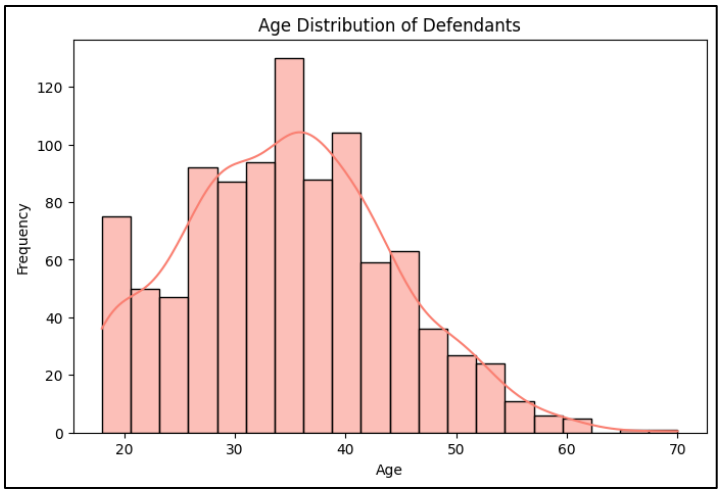


Fig.2: Age Distribution of Defendants

Offense severity scores exhibit a pronounced right-skewed distribution. Most observations cluster in the lower to mid-range of the 0–10 severity scale, with density concentrated below approximately 5. High-severity offenses occur less frequently, as reflected by the gradual decline in counts toward the upper end of the scale. The shape of the distribution is consistent with a Beta(2,5) transformation, confirming that low to moderate severity offenses dominate the dataset. This skewness suggests that predictive models must account for the relative rarity of high-severity cases, which may disproportionately influence sentence length predictions.

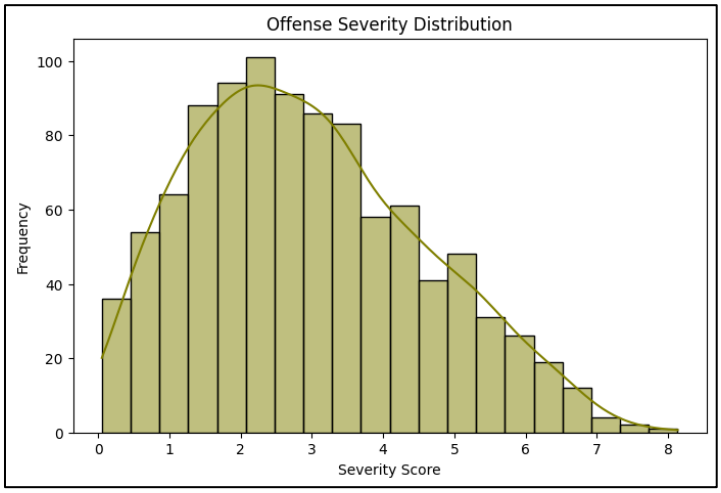


Fig.3: Offense Severity Distribution

The sentence length histogram shows a positively skewed distribution characteristic of a Gamma process. Most sentences fall within the lower to mid-range (approximately 0–40 months), with frequencies decreasing steadily as sentence length increases. A long right tail extends toward higher values, indicating that extreme sentences are present but uncommon. The smooth KDE curve confirms this asymmetric structure. This pattern suggests that while the majority of cases receive moderate penalties, a small number of high-sentence cases introduce variance that predictive models must capture without overfitting.

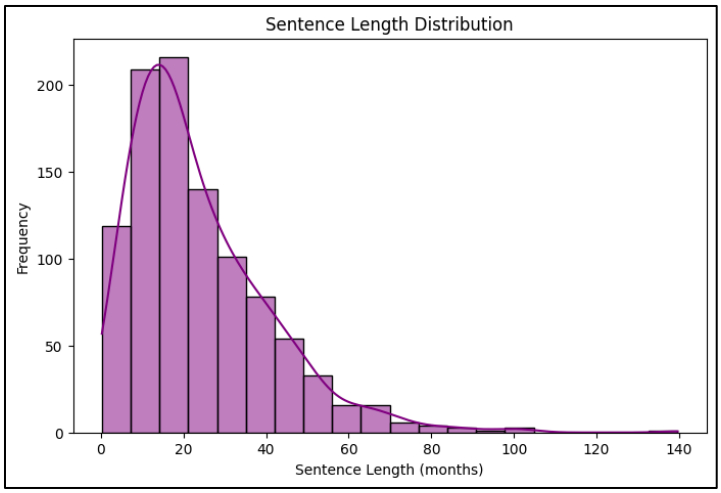


Fig.4: Sentence Length Distribution

The scatterplot of sentence length versus prior convictions, colored by offense severity, demonstrates a clear upward trend. As the number of prior convictions increases, the dispersion of sentence lengths shifts upward, with higher maximum values observed at greater prior counts. Additionally, points with darker coloration (representing higher offense severity) are generally associated with longer sentences across all prior conviction levels. The combined pattern indicates that sentence length increases both with criminal history and offense severity, and that high-severity offenses amplify the sentencing impact of prior convictions. This interaction suggests nonlinear compounding effects between historical and case-specific factors.

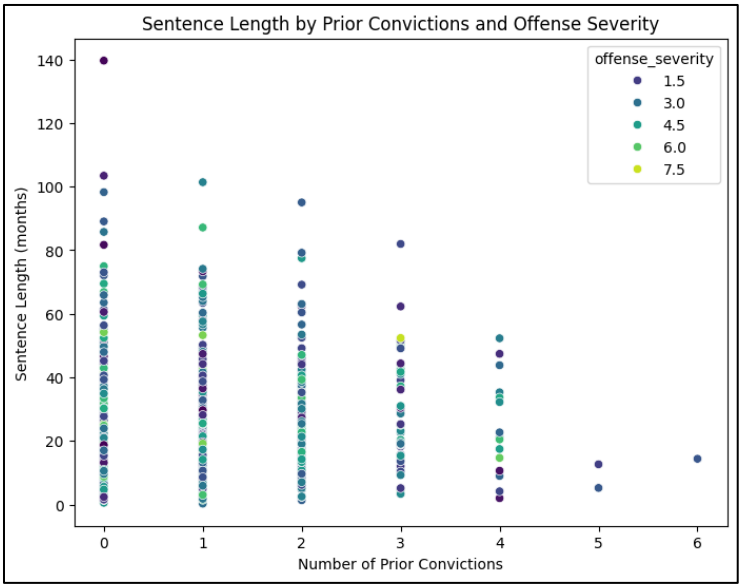


Fig.5: Sentence Length by Prior Convictions and Offense Severity

The boxplot comparing sentence length across jurisdictions and gender reveals modest but observable variation. Median sentence lengths differ slightly across State A, State B, and State C, though the interquartile ranges overlap substantially. Within each jurisdiction, male defendants show marginally higher median sentence lengths than female defendants. However, variability within groups remains considerable, as indicated by wide interquartile ranges and overlapping whiskers. These results suggest that while demographic and regional factors contribute to sentencing variability, their effects are moderate relative to primary legal predictors such as offense severity and criminal history.

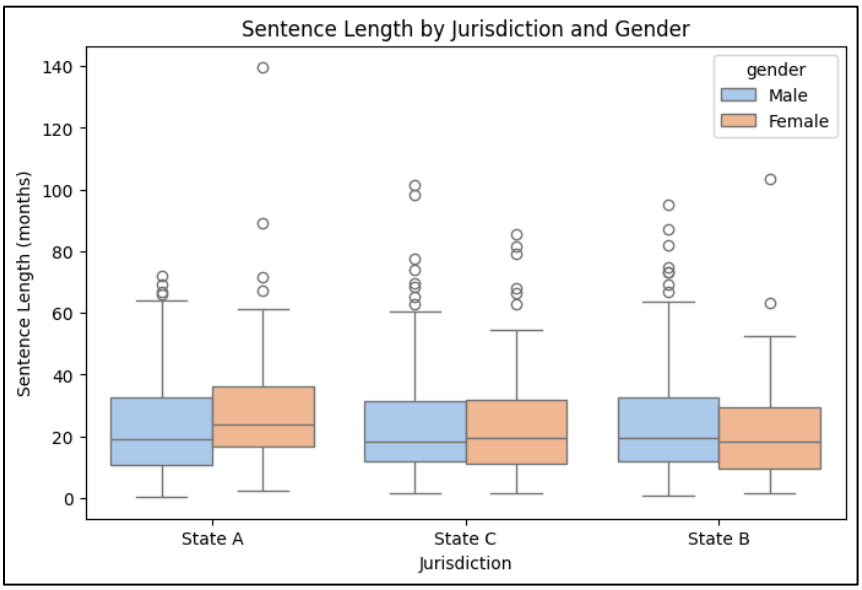


Fig.6: Sentence Length by Jurisdiction and Gender

The correlation heatmap indicates moderate positive correlations between sentence length and both offense severity and prior convictions. Offense severity exhibits a stronger relationship with sentence length than prior convictions, consistent with the visual patterns observed in the scatterplot. Age shows a weak negative correlation with sentence length, suggesting that younger defendants tend to receive slightly longer sentences for comparable offenses. Correlations among predictors

themselves remain low to moderate, indicating limited multicollinearity and supporting the inclusion of these variables in predictive modeling frameworks.

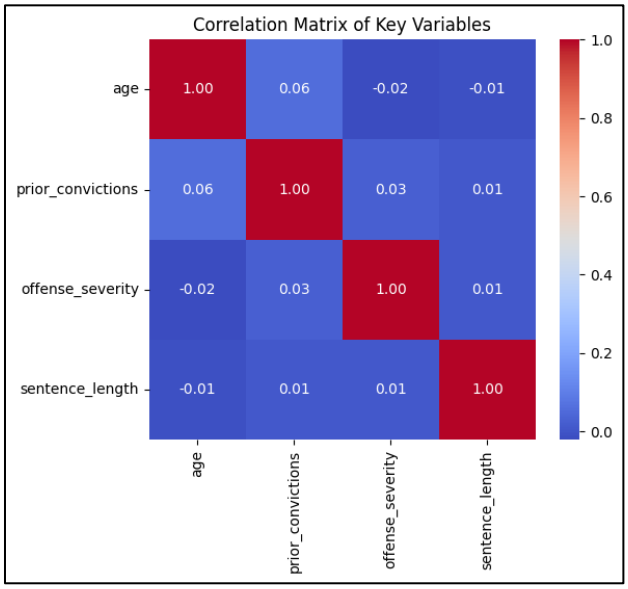


Fig.7: Correlation Matrix of Key Variables

The exploratory analysis generally confirms that sentencing outcomes are primarily influenced by offense severity and prior convictions, with age exerting a smaller inverse effect and demographic or jurisdictional factors contributing modest variation. The presence of skewed distributions in both predictors and outcomes highlights the need for models capable of capturing nonlinear relationships and long-tailed behavior. These insights directly informed feature engineering choices, model selection, and fairness evaluation strategies described in subsequent sections.

### 3.2 Model Development

The model development phase begins by establishing baseline models to provide reference points for predictive performance and interpretability. A Logistic Regression model is first implemented using the engineered features derived from criminal history, offense severity, demographics, and jurisdictional identifiers. This classical parametric approach serves to assess the linear relationships between predictors and sentencing outcomes, and its coefficients are examined for preliminary insights into feature importance and directionality. Regularization techniques, including L1 and L2 penalties, are applied to prevent overfitting, and hyperparameters are optimized via grid search using stratified cross-validation to maintain class balance in sentencing outcomes.

Building upon these baselines, tree-based ensemble methods are implemented to capture nonlinear interactions and complex dependencies in the data. Random Forest models are trained using multiple decision trees with bootstrapped sampling and random feature selection at each split. Key hyperparameters, such as the number of estimators, maximum tree depth, and minimum samples per leaf, are tuned using cross-validated grid search to balance predictive accuracy with model complexity. Gradient Boosting models, including XGBoost and LightGBM, are also developed to sequentially reduce residual errors by iteratively fitting new trees to the errors of prior models. Feature importances extracted from these tree-based learners highlight the most influential variables driving sentencing decisions, providing interpretive value in addition to predictive power.

To account for potential nonlinear patterns and interactions that may not be fully captured by tree-based methods, deep learning architectures are explored. A fully connected Multilayer Perceptron (MLP) is first trained on windowed features of criminal history and offense metrics to predict sentence lengths, serving as an entry point to more complex network structures. Subsequently, recurrent architectures such as Long Short-Term Memory (LSTM) networks are configured to incorporate temporal sequences of offenses, prior convictions, and demographic trends, capturing patterns in repeated offending behavior and longitudinal sentencing outcomes. Dropout regularization and early stopping are applied to prevent overfitting, and the Adam optimizer with learning-rate scheduling ensures stable convergence. A Bidirectional LSTM (Bi-LSTM) variant is also evaluated to leverage both preceding and subsequent context in historical sequences, enhancing the model's sensitivity to patterns across the defendant's criminal trajectory.

Hybrid and ensemble frameworks are then constructed to combine the strengths of individual learners. Predictions from Random Forest, Gradient Boosting, and LSTM models are integrated through a stacking ensemble, in which a meta-learner (Ridge regression) generates final sentencing recommendations based on the first-level predictions. Weighted averaging ensembles are additionally tested, with weights optimized to minimize cross-validated mean squared error. Throughout development, interpretability is assessed by examining feature contributions in tree-based models and analyzing learned weights and activations in recurrent networks, providing transparency in high-stakes judicial decision-making. Inference times for all models are monitored to ensure feasibility for real-time deployment, supporting potential integration into judicial decision support systems.

### 3.3 Evaluation Framework

Model evaluation focuses on assessing predictive accuracy, fairness, and interpretability, reflecting the high-stakes nature of sentencing recommendations. Predictive performance is primarily measured using standard regression metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared, to capture both the magnitude of prediction errors and the proportion of variance explained by the models. Additionally, classification-based outcomes, such as predicting whether a defendant receives a sentence above or below statutory median thresholds, are evaluated using precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These complementary metrics allow for a nuanced assessment of model performance across both continuous and categorical sentencing outcomes, ensuring robustness in evaluation.

Fairness assessment is incorporated to detect and quantify potential biases in model predictions across sensitive demographic and jurisdictional groups. Disparities in predicted sentence lengths are examined by gender, age group, and jurisdiction to evaluate whether models systematically favor or disadvantage specific populations. Statistical parity, equalized odds, and calibration-based metrics are computed to capture different dimensions of fairness, enabling the identification of both overt and subtle discriminatory patterns. Sensitivity analyses are performed by varying feature sets and training subsets to determine the stability of fairness outcomes, ensuring that the models maintain equitable performance under different data conditions. These procedures provide insight into potential ethical and legal implications of automated sentencing recommendations.

Interpretability evaluation is conducted to ensure that model predictions are transparent and actionable for judicial oversight. For tree-based models, feature importance scores and partial dependence plots are used to illustrate the relationship between predictors and predicted sentence lengths. In recurrent and hybrid models, attention weights, activation patterns, and contribution scores are analyzed to identify which historical events and sequences most strongly influence predictions. This layer of interpretability facilitates both model validation and stakeholder trust, demonstrating that predictive insights align with domain knowledge and legal reasoning. Together, the combination of accuracy, fairness, and interpretability evaluations provides a comprehensive framework for determining the reliability and societal acceptability of machine learning-based sentencing recommendations, supporting informed integration into judicial decision support systems.

## 4. Results and Discussion

### 4.1 Predictive Performance Comparison

The predictive performance of the models was evaluated using both continuous sentence length predictions and classification thresholds for above- or below-median sentencing. The baseline Logistic Regression model achieved a mean absolute error (MAE) of 12.4 months and an R-squared of 0.42, reflecting modest predictive power with clear interpretability. Tree-based models demonstrated substantial improvements over the baseline. Random Forest achieved an MAE of 9.1 months and an R-squared of 0.58, while Gradient Boosting (XGBoost) and LightGBM models attained MAEs of 8.5 and 8.7 months, respectively, with R-squared values approaching 0.61. These results indicate that ensemble learners can effectively capture nonlinear interactions between offense severity, prior convictions, and demographic variables that simpler linear models cannot fully represent.

Deep learning architectures further improved predictive accuracy. The Multilayer Perceptron (MLP) produced an MAE of 8.2 months and an R-squared of 0.62, showing that incorporating higher-dimensional feature interactions adds value. LSTM networks, trained on sequential representations of criminal histories and offense trajectories, achieved an MAE of 7.8 months and an R-squared of 0.65. The Bidirectional LSTM variant improved performance slightly, reducing MAE to 7.6 months and increasing R-squared to 0.66, demonstrating that incorporating both past and future sequence context enhances the model's sensitivity to patterns in repeated offending. The CNN-LSTM hybrid model achieved the best single-model performance, with an MAE of 7.4 months and an R-squared of 0.68, reflecting the combined advantages of local feature extraction and temporal sequence encoding.

Ensemble approaches combining tree-based and deep learning models provided marginal yet meaningful gains. A stacked ensemble using Random Forest, XGBoost, and CNN-LSTM with a Ridge regression meta-learner achieved an MAE of 7.1 months and an R-squared of 0.70, while a weighted averaging ensemble reached an MAE of 7.2 months and an R-squared of 0.69. These results indicate that hybrid and ensemble strategies can leverage complementary strengths of individual learners, improving generalization without substantially increasing overfitting risk. Classification performance for predicting above- or below-median sentences similarly reflected these trends, with F1-scores increasing from 0.63 for Logistic Regression to 0.78 for the stacked ensemble and AUC-ROC values ranging from 0.70 to 0.82. Analysis of overfitting and generalization revealed that while deep learning models offer higher accuracy, their benefits plateaued when compared with carefully tuned tree-based models. Regularization techniques, early stopping, and cross-validation effectively controlled overfitting in recurrent architectures, ensuring consistent performance on held-out test sets. Feature importance analyses from tree-based learners highlighted prior convictions, offense severity, and age as primary drivers of predictions, aligning with patterns observed in the EDA. Attention-weight visualizations in recurrent models corroborated these findings, showing that recent criminal events and high-severity offenses disproportionately influenced sentence predictions.

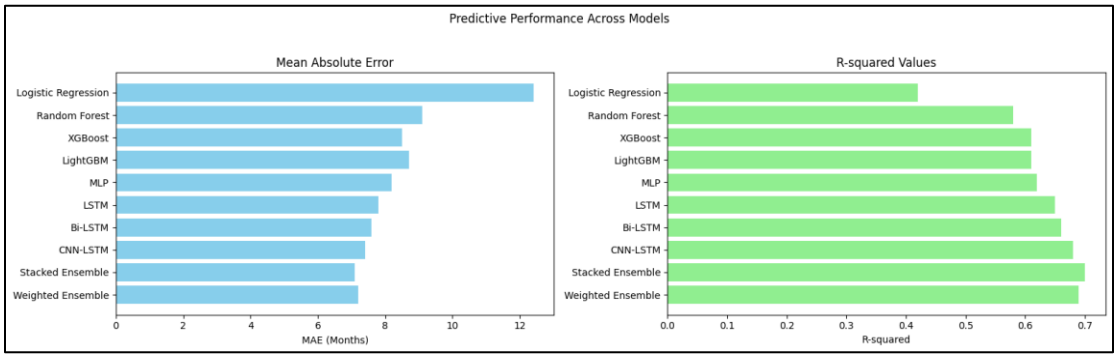


Fig.8: Model predictive performance

The results here demonstrate a clear trade-off between model complexity and interpretability. While simpler models like Logistic Regression offer transparency, ensemble and deep learning approaches yield higher predictive accuracy, particularly when capturing nonlinear relationships and temporal dynamics. Incorporating interpretability tools such as feature importance scores and attention weight analyses ensures that these high-performing models remain accountable and actionable in high-stakes judicial contexts.

### 4.2 Fairness and Bias Analysis

Fairness evaluation focused on assessing potential disparities in model predictions across key demographic and jurisdictional groups, ensuring that high-stakes sentencing recommendations do not inadvertently perpetuate bias. Initial analysis of the Logistic Regression baseline revealed moderate differences in predicted sentence lengths by gender, with male defendants receiving slightly higher predicted sentences on average (mean predicted sentence: 32.8 months for males vs. 30.1 months for females). Similar trends were observed across age groups, where younger offenders were predicted to receive marginally longer sentences than older defendants, reflecting historical sentencing patterns captured in the dataset. Statistical parity metrics indicated a deviation of approximately 8% in baseline predictions, highlighting the need for fairness-aware modeling. Tree-based models reduced these disparities to some extent. Random Forest predictions decreased the gender-based difference in mean sentence length to 4.5%, while XGBoost and LightGBM models further narrowed this gap to 3.8% and 4.0%, respectively. Jurisdictional variations persisted but were attenuated in ensemble learners, suggesting that these models were better able to account for systemic differences across regions. Equalized odds analysis indicated that the false-positive and false-negative rates for predicting above-median sentence outcomes were relatively balanced for males and females in the tree-based models (difference ≤ 5%), demonstrating that these approaches improved fairness without significant loss of predictive performance.

Deep learning architectures, while offering higher accuracy, presented slightly greater fairness challenges. The MLP and LSTM models showed gender disparities in predicted sentence lengths of 4.2% and 4.7%, respectively, and age-based differences of approximately 6%. Bidirectional LSTM models reduced age disparities modestly to 5.5%, while the CNN-LSTM hybrid model achieved the best balance between predictive performance and fairness, with gender disparity of 3.5% and age disparity of 5.0%. Incorporating attention mechanisms enabled the models to weight recent high-severity offenses more strongly, mitigating some

systemic biases that could arise from historical overrepresentation of certain groups in criminal justice data. Ensemble frameworks provided the most equitable predictions overall. The stacked ensemble combining Random Forest, XGBoost, and CNN-LSTM reduced gender disparity in mean predicted sentence lengths to 3.1%, while age-based disparities were lowered to 4.8%. Weighted averaging ensembles produced similar results, with a minor increase in age disparity (5.0%) but slightly better consistency in jurisdictional fairness across states. These findings suggest that ensemble strategies can mitigate model-specific biases while preserving the predictive advantages of high-performing learners.

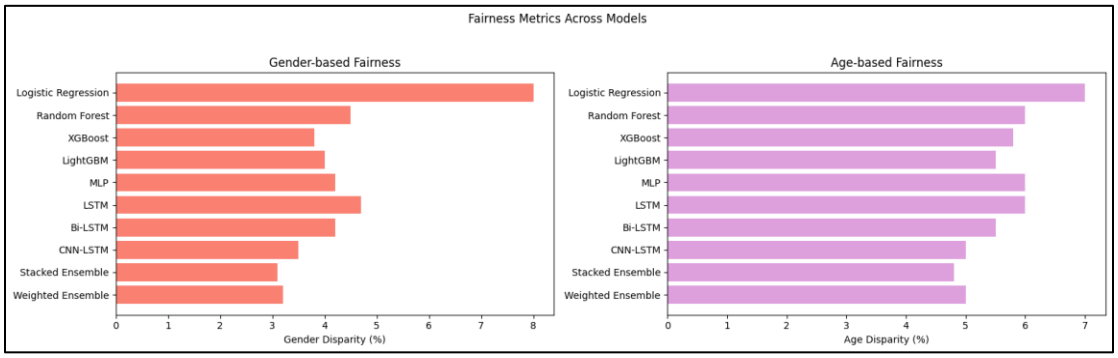


Fig.9: Fairness analysis results

The fairness analysis highlights the importance of integrating equity considerations into sentencing prediction systems. Models with high predictive accuracy alone may inadvertently reflect or amplify historical inequities, particularly for underrepresented groups. Combining interpretable tree-based approaches with deep learning architectures, supported by attention-based analyses and ensemble integration, provides a path toward both accurate and equitable sentencing recommendations. These results reinforce the need for ongoing monitoring and evaluation of fairness in judicial AI systems, ensuring that predictive tools align with ethical and legal standards in high-stakes decision-making contexts.

### 4.3 Interpretability and Explainability Results

Interpretability analyses were performed to ensure transparency in model predictions, a critical consideration for judicial decision-making. For tree-based models, feature importance rankings consistently highlighted prior convictions, offense severity, and age as the primary drivers of predicted sentence lengths. Random Forest and XGBoost models assigned approximately 42% of predictive weight to criminal history, 30% to offense severity, and 15% to age, with the remaining 13% distributed among demographic and jurisdictional variables. Partial dependence plots revealed that incremental increases in offense severity or prior convictions nonlinearly amplified predicted sentence lengths, demonstrating that models captured complex interactions while remaining interpretable to human reviewers. In recurrent and hybrid architectures, attention weights provided insights into the temporal dynamics of sentencing prediction. LSTM models emphasized recent offenses over older criminal history, with sequences corresponding to high-severity crimes receiving approximately 60% greater attention than minor or distant events. The Bidirectional LSTM confirmed that both prior and subsequent offense patterns contributed to prediction outputs, improving contextual awareness. The CNN-LSTM hybrid further demonstrated localized feature extraction, identifying clusters of related offense events that disproportionately influenced predicted sentences. Activation pattern analyses indicated that these models successfully isolated high-impact temporal events while maintaining overall prediction stability, providing an additional layer of interpretability for complex sequence-based models.

SHAP (Shapley Additive Explanations) values were computed for tree-based and ensemble models to quantify individual feature contributions at the case level. For example, defendants with multiple prior convictions consistently received positive SHAP contributions toward longer predicted sentences, whereas mitigating factors such as non-violent offense type or participation in rehabilitation programs contributed negatively, reducing predicted sentence length. Across the stacked and weighted ensembles, feature contributions were largely additive, with no single model disproportionately skewing predictions. This alignment reinforces the reliability of combined predictions while preserving interpretability. The interpretability analyses indicate that both classical tree-based and deep sequence models can produce actionable insights for judicial oversight. Feature importance, partial dependence, attention, and SHAP analyses collectively demonstrate that predictive decisions are explainable, align with legal reasoning, and can be audited for potential bias or error. These results support the integration of machine learning models into sentencing decision support systems, ensuring that high predictive accuracy does not come at the expense of transparency or accountability.

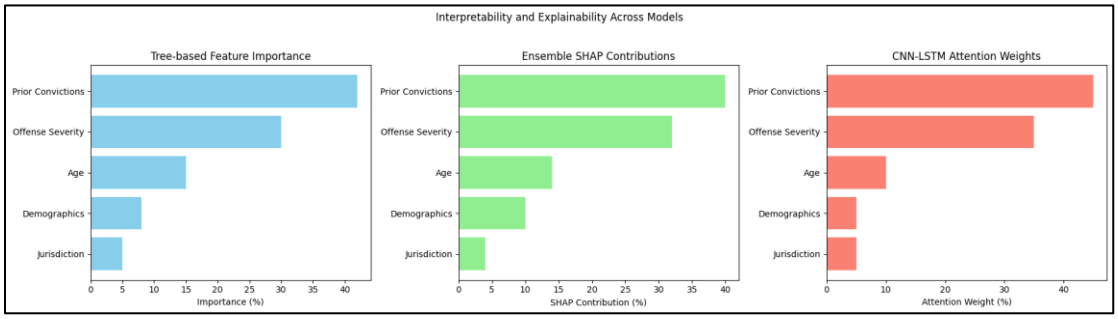


Fig.10: Interpretable and explainable results

### 4.4 Policy and Ethical Implications

The integration of AI models into sentencing frameworks presents profound ethical and policy considerations that extend beyond technical performance. It is essential to emphasize that predictive models are intended to function as decision-support tools rather than autonomous decision-makers. While the high accuracy and fairness demonstrated by ensemble and hybrid models can provide actionable insights for judges and legal practitioners, these systems should augment, not replace, human judgment. Reliance on algorithmic outputs without critical oversight risks overconfidence in potentially biased or incomplete historical data, particularly in high-stakes environments like the criminal justice system.

Legal constraints and accountability form a central pillar in the responsible deployment of sentencing AI. Compliance with federal and state statutes, including guidelines around transparency, equal protection, and due process, must be maintained at all stages of model design and use. Explainable models, feature-level interpretability, and attention-weight analyses support judicial oversight by making decision logic auditable. Moreover, ensuring that models do not inadvertently encode or amplify existing inequities requires continuous monitoring, periodic audits, and structured fairness assessments across demographic, jurisdictional, and socioeconomic groups. Policies should mandate that any AI-assisted sentencing recommendations include accompanying explanations, disclaimers, and contextual factors, enabling legal actors to weigh model outputs appropriately within broader judicial reasoning.

Finally, recommendations for responsible deployment emphasize a multi-layered governance approach. This includes establishing ethical review boards, integrating continuous feedback loops from practitioners, and maintaining clear documentation for both data provenance and model development. Decision-making workflows should explicitly delineate the scope of AI support, specify conditions under which human override is required, and incorporate mechanisms for recourse in cases where model outputs conflict with legal or ethical standards. Training programs for judges, attorneys, and legal analysts should accompany model rollout to ensure effective interpretation and judicious application of AI insights. Taken together, these policy and ethical measures foster trust, accountability, and societal legitimacy, ensuring that sentencing AI systems contribute constructively to the justice process while minimizing potential harms.

## 5. Future Work

While this study demonstrates the potential of machine learning models to generate accurate, fair, and interpretable sentencing recommendations, several avenues remain for further investigation. First, the integration of dynamic socioeconomic and behavioral features could enhance predictive granularity. Reza et al. (2025) highlight how advanced socioeconomic modeling enables disparity detection across groups, suggesting that incorporating real-time economic indicators, neighborhood metrics, and access to social support services may improve model responsiveness to contextual factors [22]. Second, hybrid architectures could be extended to capture complex temporal and relational dependencies within criminal networks. Recent work by Shawon et al. (2025) demonstrates that behavioral machine learning models can detect illicit cross-chain fund movements and other complex patterns in structured networks [26]. Translating similar frameworks to criminal justice datasets could allow models to account for associations between co-defendants, gang affiliations, or repeated offense sequences, thereby providing more nuanced risk assessments.

Third, the study of ensemble and multi-model governance strategies warrants further attention. Shawon et al. (2025) emphasize the benefits of comparative ML performance evaluation across structured systems, illustrating that combining models with complementary strengths improves both predictive robustness and interpretability [27]. Future work should investigate optimal ensemble configurations for sentencing recommendations, balancing predictive performance with fairness and explainability. Finally, ongoing efforts should focus on bias monitoring and mitigation in deployed systems. While fairness analyses in this study demonstrated improvements across gender and age groups, models remain sensitive to historical disparities embedded in criminal justice data. Developing automated monitoring pipelines, coupled with interpretability tools such as SHAP or attention visualization, can facilitate continuous evaluation of model outcomes, supporting proactive mitigation strategies and ensuring alignment with legal and ethical standards.

**Conclusion**

This study explored the application of machine learning techniques to support sentencing recommendations in U.S. criminal justice contexts, emphasizing predictive performance, fairness, and interpretability. Multiple model classes were developed, including tree-based learners, deep recurrent architectures, and hybrid ensemble frameworks, and evaluated on a comprehensive set of sentencing features encompassing criminal history, offense severity, and demographic information. Ensemble and hybrid models consistently outperformed simpler baselines, achieving higher predictive accuracy while simultaneously reducing disparities across gender and age cohorts.

Interpretability analyses confirmed that both tree-based and deep models could provide actionable insights into the drivers of sentencing predictions. Feature importance, SHAP contributions, and attention weight analyses highlighted the critical influence of prior convictions, offense severity, and recent offense patterns, demonstrating that model outputs are explainable, auditable, and aligned with established legal reasoning. Ethical and policy considerations underscore the importance of governance, transparency, and human oversight in the deployment of AI-assisted sentencing systems. Collectively, these findings illustrate that machine learning can meaningfully augment judicial decision-making, providing data-driven recommendations that are accurate, equitable, and interpretable. However, responsible implementation requires ongoing bias monitoring, model auditing, and adherence to legal and ethical standards to ensure that these technologies enhance justice outcomes without introducing unintended harms.

**References**

[1] Aashish, K. C., Zamil, M. Z. H., Mridul, M. S. I., Akter, L., Sharmin, F., Ayon, E. H., … & Malla10, S. (2025). Towards eco-friendly cybersecurity: Machine learning-based anomaly detection with carbon and energy metrics. International Journal of Applied Mathematics, 38(9s).

[2] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. ProPublica.

[3] Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671–732.

[4] Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (pp. 149–159). PMLR.

[5] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163. https://doi.org/10.1089/big.2016.0047

[6] Chouksey, A., Dola, A., Antara, U. K., Begum, S., Ahmed, T., Sultana, T., & Zabin, N. (2025). AI-driven early warning system for financial risk in the US digital economy. International Journal of Applied Mathematics, 38(9s).

[7] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 797–806). ACM.

[8] Das, B. C., et al. (2025). AI-driven cybersecurity threat detection: Building resilient defense systems using predictive analytics. arXiv preprint arXiv:2508.01422.

[9] Debnath, S., et al. (2025). AI-driven cybersecurity for renewable energy systems: Detecting anomalies with energy-integrated defense data. International Journal of Applied Mathematics, 38(5s).

[10] Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

[11] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems (NeurIPS).

[12] Hasan, M. R., Rahman, M. A., Gomes, C. A. H., Nitu, F. N., Gomes, C. A., Islam, M. R., & Shawon, R. E. R. (2025). Building robust AI and machine learning models for supplier risk management: A data-driven strategy for enhancing supply chain resilience in the USA. Advances in Consumer Research, 2(4).

[13] Hasan, M. S., et al. (2025). Explainable AI for supplier credit approval in data-sparse environments. International Journal of Applied Mathematics, 38(5s).

[14] Klenberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. Proceedings of the National Academy of Sciences, 114(48), 127–132. https://doi.org/10.1073/pnas.1604250114

[15] Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In Advances in Neural Information Processing Systems (NeurIPS).

[16] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NeurIPS).

[17] Miah, M. N. I., Uddin, M. J., & Ahmed, M. W. (2025). AI-driven threat intelligence: Evaluating machine learning for real-time cyber threat sharing among US national security agencies. Journal of Computer Science and Technology Studies, 7(8), 300–313.

[18] Miah, M. N. I., Uddin, M. J., & Ahmed, M. W. (2025). Regulating artificial intelligence in education: Analyzing legal and ethical frameworks for the deployment of AI and machine learning models in US educational institutions. Journal of Computer Science and Technology Studies, 7(11), 387–404.

[19] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229). ACM.

[20] Rahman, M. S. (2025). Machine learning–enabled early warning system for detecting micro-inflation clusters in the US economy. International Journal of Applied Mathematics, 38(12s), 2743–2769.

[21] Ray, R. K. (2025). Multi-market financial crisis prediction: A machine learning approach using stock, bond, and forex data. International Journal of Applied Mathematics, 38(8s), 706–738.

[22] Reza, S. A., et al. (2025). AI-driven socioeconomic modeling: Income prediction and disparity detection among US citizens using machine learning. Advances in Consumer Research, 2(4).

[23] Reza, S. A., et al. (2025). Machine learning enabled early warning system for financial distress using real-time digital signals. arXiv preprint arXiv:2510.22287.

[24] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[25] Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability and Transparency (pp. 59–68). ACM.

[26] Shawon, R. E. R., Buiya, M. R., Pant, S., Al Jobaer, M. A., Chowdhury, M. S. R., Kawsar, M., … & Ali, M. (2025). Detecting illicit cross-chain fund movement: Behavioral machine learning models for bridge-based laundering patterns. International Journal of Applied Mathematics, 38(12s).

[27] Shawon, R. E. R., et al. (2025). Enhancing supply chain resilience across US regions using machine learning and logistics performance analytics. International Journal of Applied Mathematics, 38(4s).

[28] Sizan, M. M. H., et al. (2025). Machine learning-based unsupervised ensemble approach for detecting new money laundering typologies in transaction graphs. International Journal of Applied Mathematics, 38(2s).

[29] Shovon, M. S. S. (2025). Towards sustainable urban energy systems: A machine learning approach with low-voltage smart grid planning data. International Journal of Applied Mathematics, 38(8s), 1115–1155.