Opening the Black Box: How Boolean AI can Support Legal Analysis

***Abstract*—In crime scene scenarios, there are various factors to consider when determining a suspect's guilt. However, the process of extracting and assessing these factors can be time-consuming, often taking years and incurring significant legal expenses. Judges are now exploring the potential of artificial intelligence techniques and machine learning computations within the justice system. Specifically, in the realm of criminal justice, these methodologies have the potential to aid in investigations and decision-making processes. Utilizing machine learning approaches can thus expedite the bureaucratic process, potentially making it more efficient. We introduce an idea of an approach that could provide fast and explainable support in the evaluation of guilt. Our approach relies on computations based on the presence or absence of 44 features describing the crime scene. Then, by a boolean function, we determined the final verdict of the legal case (only a subset of the extracted features are relevant to evaluate the guilt prediction). To demonstrate the practicality of our proposal, we conducted experiments based on 79 road homicide cases in Italy. As a consequence, the boolean evaluation was done according to Italian law principles. With our system, we reached a 83.2% accuracy rate in extracting features from the legal ruling texts and a 69.6% accuracy in guilt prediction.**

I. INTRODUCTION

Integrating Artificial Intelligence (AI) and Machine Learning (ML) within various societal domains has been transformative, particularly in fields requiring complex decision-making processes like the legal system. In recent years, Italian jurisprudence has seen a growing interest in employing ML techniques to enhance judicial efficiency [1], [2]. Traditional legal analyses often lead to lengthy proceedings due to their reliance on manual feature extraction and deliberation. Judges are exploring AI techniques to assist and speed up decisions on processes where there are multiple aspects to evaluate, such as determining guilt or innocence [1], [2].

Our work presents a novel *"boolean AI"* approach based on the workflow shown in Figure 1 that leverages AI for feature extraction combined with a law-specific rule-based function to expedite legal decision-making, thereby addressing the issues of protracted legal processes and associated costs. More specifically, our systems extracts 44 features describing crime scenes via a neural network and a decision tree, and a subsequent rule-based function consists in a boolean function evaluating a subset of the extracted features, the ones relevant to evaluate a guilt prediction according to the Italian law. This mechanism provides a transparent and logical sequence of operations, aligning with the need for justifiable and comprehensible decisions in the legal domain [3].

[†]Grazia Garzo and Stefano Ribes contributed equally to this manuscript; therefore, they have to be considered both as first author.

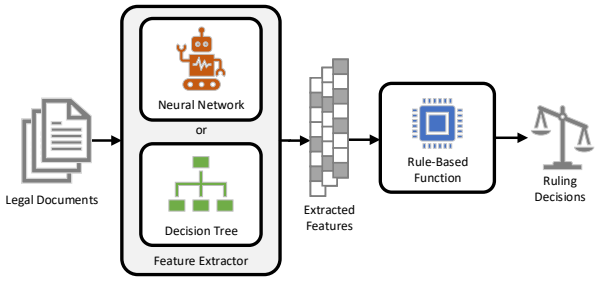

Figure 1: The overview of the proposed approach.

The synergy of the proposed approach improves ethical and legal accountability, adhering to the need for transparency in judicial proceedings. Additionally, the system's explainability facilitates human oversight, allowing legal professionals to review, understand, and potentially correct AI recommendations, thus safeguarding against errors and biases [4], [5].

This work aims to introduce a novel yet simple "boolean AI" approach that combines AI-based feature extraction with law-specific rule-based functions. This integration streamlines legal decision-making processes, addressing the issues of prolonged proceedings and associated costs in the legal domain [3]. Our approach also offers a high degree of explainability, allowing transparency in legal decision-making. We achieve this by bridging the gap between complex computational models and structured legal frameworks, promoting transparency, human oversight, and the potential correction of AI recommendations by legal professionals [4]–[7]. To demonstrate our methodology's potential in improving legal text analysis, we evaluated our system on a dataset of 79 Italian road homicide cases. The top accuracy reached is 0.832 in describing the extraction of the crime scene features from legal texts; the maximum boolean guilt prediction one is 0.696%.[1]

II. RELATED WORK

The field of decision-making AI support has garnered significant attention in recent years, with various works addressing this topic [2]. Within this domain, several AI-based techniques have been developed to aid in decision-making processes, showcasing the versatility of AI. For instance, Travaini et al. present a digital prediction model for recidivism probability among convicts [8], while Simmler et al. offer AI-based support to Swiss police in predicting crime-prone areas in Switzerland [9]. Similarly, Barnett et al. introduce the concept

[1]Anonymized code available at: https://anonymous.4open.science/r/ml4justice-boolean-ai/

of "JudicialTech" technology, which aims to enhance access to justice and potentially increase fairness in the judicial system [10]. Our approach shares similarities with these works, as we also utilize AI for decision-making in the legal context. Shaikh et al. focus instead on categorizing individuals as acquitted or convicted based on data from criminal cases in Delhi District Court murder incidents. They employ various ML models, achieving accuracy ranging from 85% to 92% and F1 scores between 86% and 92% [11]. While our study considers a similar number of cases, our methodology differs in the approach to feature extraction and decision-making. Meanwhile, Chou et al. present a methodology using the Support Vector Machine algorithm (SVM) for document classification, clustering and search to enhance law enforcement departments' efficiency in managing written criminal judgments. They achieve accuracy ranging from 73.91% to 89.49% depending on the datasets [12]. Similarly, Luo et al. evaluate different neural networks and SVM implementations to forecast criminal charges, achieving higher accuracy scores and F1 values, reaching 98.97% and 98.51%, respectively, with lower values at 42.9% and 41.16% [13]. In the context of the United Kingdom, Strickson et al. propose a digital system for predicting legal decisions or judgments within the United Kingdom. They utilize English documents and supervised learning algorithms, achieving accuracy rates between 49.6% and 69.1% [14]. In a similar approach, Mahmoudi et al. employ a transformer-based judicial algorithm, *CamemBERT*, to predict judicial decisions, with an average accuracy of 84.1% and an F1 score of 89.55% [15]. In summary, our work contributes to the growing body of research in AI-supported decision-making within the legal domain, offering a distinct approach that emphasizes explainability and transparency. This aligns with the current technological advancements in AI and ML [2], [3], enhancing the overall landscape of decision-making support in legal contexts.

III. METHODOLOGY

The dataset for this study was compiled from 79 road homicide cases in Italy. Each case is represented by a text description reporting 44 aspects describing the crime scene (*i.e.*, presence of a car, presence of a scooter, accident that occurred in the city center, *et cetera*). The relevant features for the conviction evaluation are the violation ones (reported in table **??**) evaluated together with those describing the dynamics of the accident. They are a subset of all extracted features. All the 44 features were determined based on the Italian law referring to road homicide accidents [16]. In particular, we assume features being binary, *i.e.*, either being present or not in the given legal text. Because of that, we set up the feature extraction process as a multilabel classification task.

A. Model Development

We opted for a leave-one-out cross-validation (LOO-CV) strategy to evaluate our system best. This approach ensured that every sample in our limited dataset was utilized effectively, as each case was used once as the test set, with the rest serving as the training set. Due to such dataset constraints, we focused on training a simpler model architecture rather than employing several Large Language Models (LLMs), which might require more extensive data and training for optimal performance.

Table I: Violations dependencies.

Defendant's Violation (DV)	**Victim's Violation (VV)**
Phone used while driving	Belt safety missing
Rough driving	Rough driving
Highway code violation	Highway code violation
Drugged or drunk	Drugged or drunk
Not authorized vehicle	Not authorized vehicle
Speeding	Speeding
No driver's license	No driver's license

Our research employed two distinct feature extractor models: a Multi-Layer Perceptron (MLP) based model and a Sci-kit learn pipeline model integrating TF-IDF vectorization and a multilabel group of Decision Trees (DTs). The MLP model architecture was designed as a text classifier using PyTorch and PyTorch Lightning, consisting of linear layers with dropout regularization to prevent overfitting. A binary cross-entropy loss function was used to train the model for multilabel classification. For the DT-based pipeline, instead, we leveraged the Optuna optimization framework to isolate the best hyperparameters for subsequent training in the aforementioned LOO-CV setting. Considered hyperparameters included, among others: n-gram range, maximum tree depth, training criterion, *et cetera*. Please note that the models trained during hyperparameter tuning were discarded, *i.e.*, they were not used as pre-trained models during LOO-CV.

B. Training Procedure

In each iteration of LOO-CV, one case was used as the test set, while the remaining cases constituted the training set. To handle the imbalanced nature of the dataset, we applied a custom oversampling technique to ensure that minority classes were adequately represented during training. This step was crucial to prevent the model from being biased towards the majority classes most frequently appearing in the legal texts. Overall, LOO-CV ensured that every case was used for testing exactly once, providing a robust evaluation of the model's performance across the entire dataset.

C. Model Evaluation

The model's performance was evaluated using various metrics (reported in Table III), including accuracy, F1 score, precision, and recall. To provide a comprehensive assessment, we employed different reduction methods for these metrics, such as micro, macro (a "per-label" score), and weighted (like macro, but averaged based on labels' supports) averages, allowing us to capture distinct aspects of model performance across labels. The rates refer to values given by the average of averages, as reported in equations 1 and 2.

$$Y_j = \frac{X_1 + X_2 + \ldots + X_{44}}{44} \tag{1}$$

Table II: Look Up Table to determine the guilt response.

FO	AO	DV	VV	Guilt
0	0	0	0	0
0	0	0	1	0
0	0	1	0	0
0	0	1	1	$*$
0	1	0	0	$-$
0	1	0	1	$-$
0	1	1	0	$-$
0	1	1	1	$-$
1	0	0	0	0
1	0	0	1	0
1	0	1	0	1
1	0	1	1	$*$
1	1	0	0	0
1	1	0	1	0
1	1	1	0	1
1	1	1	1	1

$$Accuracy = \frac{Y_1 + Y_2 + \ldots + Y_{79}}{79} \tag{2}$$

When considering accuracy, X_i refers to the rate in the detection of the i-th feature over the selected 44; Y_j refers to the average rate of the j-th sample over the 79 legal cases.

D. Rule-Based Conviction Classification

The conviction classification algorithm assesses potential culpability in road homicide cases. It evaluates key factors derived from the features extracted from the preceding ML model: Foreseeable Obstacle (FO), Avoidable Obstacle (AO), Defendant's Violation (DV), and Victim's Violation (VV). In particular, FO and AO are two predicted labels by the ML model, whereas DV (respectively, VV) is evaluated as true if any predicted feature corresponding to the defendant is true (respectively, to the victim). For example, in case of safety violations or substance abuse by the defendant, DV will be set to true. The boolean function's core decision-making logic is reported in Table II (where 1 means $True$; 0 corresponds to $False$). DVs, VVs, and the II are based on the rules of the Italian law in road homicide evaluations [16]. Algorithm 1 includes the pseudo-code implementing Table II. It is worth mentioning that it is impossible to have a scenario where the obstacle is avoidable but not foreseeable (the rows with $Guilt$ equal to $-$ in Table II). Furthermore, according to the Italian law, and referencing Table **??**, if the defendant committed more or the same number of violations compared to the victim, the defendant is condemned, as reported in the rows with $Guilt$ set to $*$ in Table II.

Algorithm 1 Conviction classification algorithm.

Require: $pred$ ▷ AI model output

$FO \leftarrow pred[FO]$
$AO \leftarrow pred[AO]$
$DV \leftarrow \boldsymbol{OR_{reduce}}(pred[\text{defendant features}])$
$VV \leftarrow \boldsymbol{OR_{reduce}}(pred[\text{victim features}])$
$defViol \leftarrow \boldsymbol{COUNT}(pred[\text{defendant features}])$
$victViol \leftarrow \boldsymbol{COUNT}(pred[\text{victim features}])$
$defConvicted \leftarrow false$
if not FO **and** AO **then**
    $defConvicted \leftarrow undecided$ ▷ Invalid scenario
else if DV **and** VV **then**
    $defConvicted \leftarrow defViol \geq victViol$
else if FO **and** DV **then**
    $defConvicted \leftarrow true$
end if

IV. RESULTS AND DISCUSSION

A. Performance on Features Extraction

Table III reports the performance metrics of the designed feature extractor models. In predicting the relevant aspects of legal texts, the MLP-based model showed an overall high Accuracy of 0.877 in the micro reduction setting, indicating a strong ability to classify common combinations of text features correctly. However, this high accuracy contrasts with a relatively low F1 score of 0.507, suggesting a disparity between the model's precision and recall. This is further underscored in the macro, *i.e.*, label-wise, reduction setting, where the F1 Score drops to 0.085, indicating a significant imbalance in the model's performance across different classes. The precision and recall metrics in the micro setting (0.725 and 0.390, respectively) further confirm this imbalance. The ROC-AUC score of 0.789 in the micro setting reflects a decent capability of the model to distinguish between groups of classes, *i.e.*, legal case scenarios. However, in the macro setting, the ROC-AUC score drops to 0.406, highlighting limitations in its ability to handle class imbalances effectively.

In contrast, the DT-based model demonstrates more balanced performance across metrics, albeit with generally lower scores than the MLP in the micro setting. It achieved an accuracy of 0.832 and an F1 score of 0.500, showing a more balanced precision-recall trade-off compared to the MLP. The precision and recall values are closer in the micro setting (0.483 and 0.520, respectively), indicating a more consistent performance across classes. The ROC-AUC score of 0.706 reflects a competent, though not outstanding, capacity in distinguishing between legal cases. In the macro and weighted settings, the DT-based model shows a notable decrease in performance metrics, with F1 scores of 0.207 and 0.505, respectively, suggesting challenges in managing class imbalances.

Overall, the MLP model excels in accuracy but struggles with the class imbalance and nuances of the legal texts, as shown by its lower macro and weighted scores. This suggests that while the model is proficient in identifying the majority labels, *i.e.*, frequent feature patterns in the texts, it may not perform as well on less-represented aspects of the text data. On the other hand, the DT-based model offers a more balanced performance across classes but does not reach the high accuracy levels of the MLP in the micro setting.

B. Performance on Ruling Decision Classification

The comparison of feature extraction methods using MLP-based and DT-based for the ruling decision algorithm shows distinct outcomes. The classification based on the features extracted by the MLP model, despite having the highest accuracy (0.785) and F1 score (0.879), exhibits a significant

Table III: Performance metrics of the feature extraction models.

Feature Extractor	Reduction type	Accuracy	F1 Score	ROC-AUC	Precision	Recall
MLP	micro	0.877	0.507	0.789	0.725	0.390
MLP	macro	0.877	0.085	0.406	0.126	0.086
MLP	weighted	0.780	0.366	0.485	0.430	0.390
DT	micro	0.832	0.500	0.706	0.483	0.520
DT	macro	0.832	0.207	0.519	0.202	0.216
DT	weighted	0.719	0.505	0.561	0.497	0.520

limitation as indicated by its ROC-AUC score of 0.500. The classification score on the features extracted by the DT-based model, while having a lower accuracy and F1 score (0.696 and 0.810, respectively), shows a somewhat better ability to distinguish between classes, with a ROC-AUC of 0.529. Given such features extraction performances, with the boolean function based on Table II and Algorithm 1, we reached a 69.6% accuracy in the guilt prediction task. The boolean function is based on the Italian relevant features for road homicides, reported in Table **??**, which are a subset of our 44 extracted elements.

C. Discussion

The application of our boolean AI system to legal text classification and ruling decision classification within the Italian legal context shows limitations. The feature extraction models, particularly the MLP, exhibit high accuracy in identifying common feature patterns in legal texts. However, this is contrasted by a notable struggle with class imbalances and nuances of the Italian legal language, as evidenced by lower F1 scores and performance drops in macro settings. This disparity suggests a proficiency in detecting frequent features but a deficiency in capturing legal texts' less common or intricate aspects.

The limitations of our approach are further compounded by the challenges posed by the complexity of the Italian legal language and the limitations of the available dataset. Legal terminology's specialized and varied nature likely contributes to the observed difficulties in model performance, particularly in accurately capturing the subtleties and less frequent patterns in legal texts. Despite these challenges, we believe in the practical potential of our system in automating aspects of legal decision-making, as it significantly streamlines the analysis of legal documents, reducing the manual effort required in processing complex legal texts.

V. CONCLUSION

We applied AI to analyze road homicide cases in Italy, employing MLP-based and DT-based features extraction models. While the MLP excelled in scenarios with less class imbalance, both models faced challenges in handling class imbalances. The DT's interpretability offers promise in contexts where decision explanation is crucial. Our research contributes to the AI field in legal decision-making, highlighting the need for models that are not only accurate but also fair, balanced, and transparent in legal settings. Our experiments were based on 79 road homicide cases in Italy. We reached 83.2% accuracy score in the extraction of features describing the crime scene from the legal ruling texts and a 69.6% accuracy in guilt prediction based on the previous extraction process. Based on our results, we believe that our approach could become a valuable tool in supporting juridical figures and lawyers in the legal processes and juridical evaluation.

REFERENCES

[1] O. A. Alcántara Francia, M. Nunez-del Prado, and H. Alatrista-Salas, "Survey of text mining techniques applied to judicial decisions prediction," *Applied Sciences*, vol. 12, no. 20, p. 10200, 2022.

[2] J. Cui, X. Shen, and S. Wen, "A survey on legal judgment prediction: Datasets, metrics, models and challenges," *IEEE Access*, 2023.

[3] C. nazionale delle ricerche (Italia). Istituto per la documentazione giuridica and R. Nannucci, *Lineamenti di informatica giuridica: teoria, metodi, applicazioni*. Ed. scientifiche italiane, 2002.

[4] M. Virzi *et al.*, "Cultura della ragionevolezza dei tempi processuali, garanzie cedu e rimedi interni: Il caso italiano tra esperienze e prospettive di riforma," 2022.

[5] G. L. Gatta *et al.*, "Prescrizione del reato e lentezza del processo: male non cura male," *SISTEMA PENALE*, 2019.

[6] UNESCO, "Unesco: Building peace in the minds of men and women," 2023. Online; accessed on December 19th.

[7] UNESCO, "Ai and the rule of law: Capacity building for judicial systems," 2024. Online; accessed on January 4th.

[8] G. V. Travaini, F. Pacchioni, S. Bellumore, M. Bosia, and F. De Micco, "Machine learning and criminal justice: A systematic review of advanced methodology for recidivism risk prediction," *International journal of environmental research and public health*, vol. 19, no. 17, p. 10594, 2022.

[9] M. Simmler, S. Brunner, G. Canova, and K. Schedler, "Smart criminal justice: exploring the use of algorithms in the swiss criminal justice system," *Artificial Intelligence and Law*, vol. 31, no. 2, pp. 213–237, 2023.

[10] J. Barnett, P. Treleaven, F. I. Lederer, N. Vermeys, and J. Zeleznikow, "Judicialtech supporting justice," *Available at SSRN*, 2023.

[11] R. A. Shaikh, T. P. Sahu, and V. Anand, "Predicting outcomes of legal cases based on legal factors using classifiers," *Procedia Computer Science*, vol. 167, pp. 2393–2402, 2020.

[12] S. Chou and T.-P. Hsing, "Text mining technique for chinese written judgment of criminal case," in *Intelligence and Security Informatics: Pacific Asia Workshop, PAISI 2010, Hyderabad, India, June 21, 2010. Proceedings*, pp. 113–125, Springer, 2010.

[13] B. Luo, Y. Feng, J. Xu, X. Zhang, and D. Zhao, "Learning to predict charges for criminal cases with legal basis," *arXiv preprint arXiv:1707.09168*, 2017.

[14] B. Strickson and B. De La Iglesia, "Legal judgement prediction for uk courts," in *Proceedings of the 3rd International Conference on Information Science and Systems*, pp. 204–209, 2020.

[15] S. A. Mahmoudi, C. Condevaux, B. Mathis, G. Zambrano, and S. Mussard, "Ner sur décisions judiciaires françaises: Camembert judiciaire ou méthode ensembliste?," in *Extraction et Gestion des connaissances EGC'2022*, 2022.

[16] P. M. Sabella *et al.*, "Art. 589-bis. omicidio stradale," in *Codice Penale: Rassegna di Giurisprudenza e di Dottrina.*, pp. 172–186, Giuffrè Francis Lefebvre, 2022.