Human Performance on the AI Legal Case Verdict Classification Task

Jack MUMFORD [a,1], Katie ATKINSON [a] and Trevor BENCH-CAPON [a]
[a] *Department of Computer Science, University of Liverpool, UK*

Abstract. We report a study undertaken to analyse human performance on the verdict classification task. Several approaches have addressed this task with outcomes compared against the outcomes from actual legal cases. Results vary and we investigate *how* classification is done by humans. A key finding is that fact descriptions alone are insufficient for accurate classification, independent of legal background.

Keywords. Legal verdict classification, Human-machine benchmarking

1. Introduction

Legal judgement prediction has featured in much recent AI and Law research, with a variety of techniques being used. Symbolic approaches model expert knowledge deployed in reasoning about legal cases [1] and many works have applied machine learning (ML) techniques to the task, e.g. [2]. The effectiveness of different approaches is evaluated according to their accuracy in matching actual outcomes of past cases. Explaining the outcomes is also a key concern for deployment of these tools. A significant body of work has emerged within the AI literature that addresses the problem of classifying the outcome of a case from a description of its facts. However, there appears to be no work asking the question of how well humans perform the task, which is the focus of this paper. A key domain that has emerged as a testbed for AI-based legal prediction and classification is the European Convention on Human Rights (ECHR). We report on a significant study undertaken to establish the level of performance of humans in classifying the verdict of cases under Article 6 (the right to a fair trial) of the ECHR. The main motivations for undertaking this study have been:

1. To establish human benchmark performance for comparison against AI tools operating on the classification task;
2. To evaluate the effect of different setups on performance at the classification task – the level of legal experience of participants, access of participants to a domain knowledge model, zero-shot vs few-shot, and reading the circumstances alone vs circumstances plus the relevant legal framework;
3. To gain insights into the process of effectively deriving outcomes of ECHR cases as a computation task.

[1] Corresponding Author: Jack Mumford, email: jack.mumford@liverpool.ac.uk

The remainder of the paper is structured as follows. Section 2 provides a brief overview of the literature on classifying legal case verdicts. Section 3 describes the setup of the human study that we ran, involving participants who determined the outcome of an ECHR legal case based on a description of the facts of the case. Section 4 presents results providing a detailed analysis of task performance. A discussion of the implications of the outcomes of the study is given in Section 5, along with closing summary remarks.

2. Classifying Legal Case Verdicts in AI and Law

Reasoning about legal cases has been a staple of AI and Law research for decades (e.g. [3]), but with the advances in Machine Learning and widespread availability of large datasets there has been an upsurge of interest in developing AI models for classifying outcomes of legal cases. The European Court of Human Rights (ECtHR) has served as a popular testbed for these tools, e.g. [2]. Many other works on automating judicial decision-making followed, both using ECtHR cases and other legal datasets. A comprehensive survey of this work is given in [4].

In [4] Medvedeva *et al.* distinguish three tasks covered by this body of work: 1) *outcome identification* (identifying the verdict in the full text of the published judgements); 2) *outcome categorisation* (categorising documents based on the outcome); 3) outcome forecasting (predicting future decisions of a particular court). Medvedeva *et al.* state that task 1 often does not require ML: keyword search can be sufficient. For task 2, the input data is based on decisions *already* made. Task 3 uses textual information about the case that is available *before* the verdict to predict the outcome. Task 3 is generally considered the hardest, since post-decision material, including the representation of the facts, may potentially contain clues as to the verdict, although this has not been proven. Much related research [2,5,6] has focused on task 2, using post-decision material with references to the outcome removed, as a practical way to simulate the conditions of task 3.

The study reported here concerns human performance at task 2 – outcome categorisation. We follow the common approach in existing literature by using post-decision material. This approach allows us to attribute any categorising inaccuracies more confidently to the participants' judgment rather than to gaps in the information provided. To evaluate the performance of machines, we need knowledge of how humans perform. The study reported in this paper establishes such a benchmark and gives insight into *how* participants perform the task and which aspects of it are the most challenging.

3. Human Study Setup

In this section we describe our study domain, the participant groups and the classification task undertaken[2]. Article 6 pertains to the **right to a fair trial**, and is often selected for training and testing of AI systems. This is largely due to the procedural nature of the article and relative abundance of data; more Article 6 cases are available on HUDOC[3] than for any other article. The reasoning of the court when deciding alleged violations of Article 6 was previously expressed in the form of an ADM (Angelic Domain Model

[2] Our study was supported by Research England under the Policy Support Funding stream.

[3] The open access ECtHR database: `https://hudoc.echr.coe.int`

[7]) knowledge model [8] that captures the legal argument from the discussion of the important legal factors of a given case, to resolution of the key issues and final outcome[4].

Each case summary is structured, with different sections delivering specific information regarding the case. Intuitively, the section that is of relevance to this paper is denoted as **THE FACTS** of the case. This section is further divided into two subsections, which we identify as the **circumstances of the case** and the **relevant legal framework**. The first of these subsections provides essential events and personal details relevant to the case, whereas the second presents the law deemed relevant for its resolution. We used regular expressions[5] and restricted to cases available in English and decided from 2015 onward (to mitigate the risk of confusion arising from changes to the law over time), to extract the two relevant subsections for each case, to form our dataset.

Intuitively, we would expect people with higher levels of legal domain experience to perform better at reliably classifying legal case outcomes correctly. By establishing benchmarks for various backgrounds, we can more accurately determine the relative performance of AI systems designed for legal case verdict classification. Consequently, we recruited participants from three student groups (all final-year students), based on their domain experience of Article 6 of the ECHR: **Weak** – computer science students with no formal legal background; **Moderate** – law students with no formal study of any module focused on the ECHR/ECtHR; **Strong** – law students with either formal study of an ECHR/ECtHR focused module, or experience on a previous research project focused on reading and interpreting cases pertaining to Article 6 of the ECHR. We further split participants to observe the effect of providing an ADM of Article 6 on participant confidence and classification performance. Approximately half of the participants were provided with the model, with the others forming the control group. We anticipated that those participants provided with the model, would have higher overall confidence in their responses. In total, 41 students were recruited, with 19 given the ADM and 22 placed into the control group. Of the 41 students, 11 were in the **Weak** group, 20 in the **Moderate** group, and 10 in the **Strong** group. Participants were paid to complete 8 hours work on the project and they were provided with 2 hours of training. The variance in the sizes of the groups is partially attributable to the late withdrawal of some individuals, and the greater ease in recruiting for the **Moderate** group.

Each participant was required to read the text from **THE FACTS** sections of ECtHR cases relating to an alleged violation of Article 6 of the ECHR. Specifically, participants would first read the subsection **circumstances of the case** and classify the outcome of the case (violation or no-violation), before reading the subsection **relevant legal framework** and making another classification based on the additional information (violation or no-violation). The distribution of verdicts was approximately balanced between violations and no violations in periodical sets of twenty cases arranged in ascending text length of **THE FACTS** section. Participants were trained on day 1, and worked on the classification tasks from days 2 – 5. Those groups provided with the ADM of Article 6 were given time during training to read and understand the content of the model and also completed two related tests[6]. Participants were employed on a zero-shot task on days 2 and 3, with no exposure to correct case verdicts. Following this, participants were pro-

[4] Access the full model at `https://github.com/jamumford/Human_Legal_Verdict_Prediction`

[5] Code available at `https://github.com/jamumford/Human_Legal_Verdict_Prediction`

[6] The quizzes and final debrief survey are now closed for responses, but can be viewed at `https://github.com/jamumford/Human_Legal_Verdict_Prediction`

Table 1. Summary of classification performance across all participant groups, where **NM** (resp. **WM**) indicates the no model (resp. with model) groups. Results in brackets indicate the standard deviations for the macro metrics, where **Productivity** indicates the mean number of classifications per participant. Productivity results are reported to 3sf, all other results are reported to 3dp.

	Micro Acc	**Macro Acc**	**Micro MCC**	**Macro MCC**	**Productivity**
Overall	0.511	0.504 (0.069)	0.091	0.079 (0.078)	88.6 (30.8)
Zero-shot	0.507	0.498 (0.073)	0.047	0.039 (0.087)	**47.3** (18.1)
Few-shot	**0.511**	**0.505** (0.127)	**0.128**	**0.130** (0.152)	40.1 (16.0)
Weak	0.491	0.481 (0.084)	0.057	0.046 (0.084)	77.5 (24.0)
Moderate	**0.532**	**0.532** (0.050)	**0.115**	**0.108** (0.059)	**97.8** (33.3)
Strong	0.483	0.475 (0.066)	0.070	0.056 (0.090)	82.4 (29.0)
NM	0.509	0.496 (0.067)	0.074	0.061 (0.072)	76.5 (24.1)
WM	**0.514**	**0.511** (0.071)	**0.102**	**0.095** (0.082)	**99.0** (32.5)
Circumstances	0.499	0.492 (0.074)	0.080	0.068 (0.079)	**44.3** (15.4)
Relevant Legal	**0.523**	**0.517** (0.077)	**0.102**	**0.093** (0.086)	**44.3** (15.4)

vided with eight practice cases that included the actual verdicts. As such, on days 4 and 5, participants were employed on a few shot learning task. When the few shot classification task had concluded, all participants completed a debrief survey to report their confidence in their knowledge and performance across the duration of study.

4. Results

In this section we present the main results of participant performance on the classification task. Our chosen metrics for performance are accuracy, MCC score, and productivity. Accuracy is included as a staple metric, but the distribution of verdicts in the dataset is not fully balanced, with 60.0% (1dp) of reviewed cases pertaining to violation verdicts (note that the real distribution for court decisions beyond the dataset is even more skewed towards violations). Hence, we focus on the MCC score as the best indicator of participant classification performance at judging the verdict outcomes in accordance with the underlying distribution. Productivity indicates the number of classifications made per participant (two per case, one for the **circumstances** and one for the **relevant legal framework**), during the eight hours of dedicated classification work. Table 1 shows the returns for group performance (micro returns use a combined confusion matrix for all participants, whereas macro preserves an individual confusion matrix for each participant) and Table 2 provides correlation tests between potentially associated variables (focusing on mean MCC as the classification performance indicator).

We can produce the following principal observations. **(Obs 1)** Overall mean participant classification performance (Table 1) is approximately equivalent to random classifier output. **(Obs 2)** There exists a slight positive (SR coefficient = 0.222) but not statistically significant (SR p-value = 0.163) correlation between participant confidence and classification performance. **(Obs 3)** Participants achieved statistically significant (MW p-value = 0.012) higher classification performance at the few-shot task compared to the zero-shot task, but there is no correlation between zero-shot and few-shot performance (SR coefficient = 0.033). **(Obs 4)** There is a statistically significant (SR p-value = 0.000) positive correlation (SR coefficient = 0.725) between participant classification performance following assessment of the **circumstances** and following assessment of the **rel-**

Table 2. Correlation statistical significance tests. Results reported to 3 decimal places.

	SR coefficient	SR p-value	MW p-value
Confidence vs Performance	0.222	0.163	n/a
Zero Shot vs Few-shot	0.033	0.843	0.012
Circumstances vs Relevant Legal	0.725	0.000	0.205
Quiz vs Performance	0.056	0.806	n/a
Productivity vs Performance	0.666	0.000	n/a

evant legal framework, but no statistically significant difference in classification performance between the two distributions (MW p-value = 0.205). **(Obs 5)** There are no statistically significant increases in classification performance associated with any other increases in participant knowledge (domain experience: best performance is by the **moderate group**, **NM** vs **WM** (MW p-value = 0.255). **(Obs 6)** There exists a statistically significant (SR p-value = 0.000) positive correlation (SR coefficient = 0.666) between participant productivity and classification performance. **(Obs 7)** There exists a statistically significant (MW p-value = 0.017) increase in productivity associated with having access to the ADM (**WM** vs **NM**).

5. Discussion and Summary Remarks

In this section, we set the main contributions against the four motivations outlined in Section 1, before discussing future research directions and limitations of the study.

(1) Human benchmark performance: We found that the overall human performance in this task closely resembled that of a random classifier, with an approximate mean accuracy of 0.5 and MCC score of 0.0.

(2) Effect of different setups: For most setups, including domain experience and knowledge of the ADM content, we found no statistically significant effects on classification performance, suggesting a limited effectiveness of university education for training law students to reconcile legal case descriptions into case outcomes. Whilst few-shot classification performance was statistically significantly higher than zero-shot performance, mean scores for accuracy and MCC were very low for both. Confidence, as indicated by participants, did not serve as a strong predictor of classification performance. Our analysis revealed two clear positive correlations: between classification performance after reading the case **circumstances** and after reading the **relevant legal framework**; and between productivity and overall classification performance. Intriguingly, participants provided with knowledge from the ADM exhibited increased productivity.

(3) Understanding how to effectively perform the classification task: Our study raises questions about the feasibility of classifying legal outcomes solely from descriptions of facts, as has been dominant in prior ML approaches to the task. To enhance accuracy, future research should consider incorporating explicit references to other cases (particularly leading cases that frequently form the reference basis for judgements) and temporal context, for establishing references to key precedents as a crucial aspect for determining verdicts. The use of advanced information/document retrieval NLP techniques would be well suited to implement these measures within AI systems designed for the classification task.

Future research: (i) Comparative analysis of state-of-the-art AI systems, such as BERT-based [9] systems designed for legal analysis [5,6], on the identical zero-shot and few-shot datasets to indicate the relative ability of these systems against the human benchmark established in this paper. **(ii)** Forecasting legal outcomes by analysing communicated case summaries (available before the case is decided) of concluded cases. **(iii)** Exploration of other legal domains may uncover domain-specific variations in performance and contribute to a more comprehensive understanding of legal decision-making.

Limitations: **(i)** Superior few-shot performance may be caused by prior experience on the zero-shot task. However, there was no statistically significant correlation between the classification performance of the two tasks. **(ii)** Scheduling of the study may have resulted in a participant sample that was not reflective of the wider student population. **(iii)** True expert knowledge could lead to substantial performance improvements. However, literature from the social sciences suggests that judges themselves do not reach sound verdicts as a function of the facts alone [10,11,12,13].

Summary: We have presented the findings of a study that involved the recruitment of students with varying legal domain experience, for the purpose of classifying verdicts related to Article 6 of the ECHR, solely from the factual descriptions of the cases. We found that mean human performance closely resembled randomness, and was unaffected by domain knowledge, underscoring the challenge of this task. Our results suggest that to enhance classification effectiveness, explicit references to other cases and temporal context should be considered, and associated advanced information retrieval techniques should be explored for implementation in AI systems. Our study thus provides valuable insights and future research directions in the domain of AI-based legal decision-making.

References

[1] Collenette J, Atkinson K, Bench-Capon T. Explainable AI tools for legal reasoning about cases: A study on the European Court of Human Rights. Artificial Intelligence. 2023;317:103861.

[2] Aletras N, Tsarapatsanis D, Preoţiuc-Pietro D, Lampos V. Predicting judicial decisions of the ECHR: A natural language processing perspective. PeerJ Computer Science. 2016;2:e93.

[3] Bench-Capon T. HYPO'S legacy: introduction to the virtual special issue. AI and Law. 2017;25(2):205-50.

[4] Medvedeva M, Wieling M, Vols M. Rethinking the field of automatic prediction of court decisions. AI and Law. 2023;31(1):195-212.

[5] Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I. LEGAL-BERT: The muppets straight out of law school. arXiv preprint arXiv:201002559. 2020.

[6] Mumford J, Atkinson K, Bench-Capon T. Reasoning with Legal Cases: A Hybrid ADF-ML Approach. In: Proceedings of JURIX 2022; 2022. p. 93-102.

[7] Atkinson K, Bench-Capon T. ANGELIC II: An Improved Methodology for Representing Legal Domain Knowledge. In: Proceedings of the 19th ICAIL; 2023. p. 12-21.

[8] Mumford J, Atkinson K, Bench-Capon T. Combining a Legal Knowledge Model with Machine Learning for Reasoning with Legal Cases. In: Proceedings of the 19th ICAIL; 2023. p. 167-76.

[9] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.

[10] Guthrie C, Rachlinski JJ, Wistrich AJ. Blinking on the bench: How judges decide cases. Cornell L Rev. 2007;93:1.

[11] Kahneman D, Klein G. Conditions for intuitive expertise: a failure to disagree. American psychologist. 2009;64(6):515.

[12] Danziger S, Levav J, Avnaim-Pesso L. Extraneous factors in judicial decisions. Proceedings of the National Academy of Sciences. 2011;108(17):6889-92.

[13] Wistrich AJ, Rachlinski JJ, Guthrie C. Heart versus head: Do judges follow the law of follow their feelings. Tex L Rev. 2014;93:855.