Legal Judgment Prediction: If You Are Going to Do It, Do It Right

Masha Medvedeva
eLaw Center for Law and Digital Technologies
& Department of Business Studies,
Faculty of Law,
Leiden University, the Netherlands
m.medvedeva@law.leidenuniv.nl

Pauline McBride
COHUBICOL,
Law, Science, Technology
and Society (LSTS),
Vrije Universiteit Brussel, Belgium
pauline.mcbride@vub.be

Abstract

The field of Legal Judgment Prediction (LJP) has witnessed significant growth in the past decade, with over 100 papers published in the past three years alone. Our comprehensive survey of over 150 papers reveals a stark reality: only ∼7% of published papers are doing what they set out to do - predict court decisions. We delve into the reasons behind the flawed and unreliable nature of the remaining experiments, emphasising their limited utility in the legal domain. We examine the distinctions between predicting court decisions and the practices of legal professionals in their daily work. We explore how a lack of attention to the identity and needs of end-users has fostered the misconception that LJP is a near-solved challenge suitable for practical application, and contributed to the surge in academic research in the field. To address these issues, we examine three different dimensions of 'doing LJP right': using data appropriate for the task; tackling explainability; and adopting an application-centric approach to model reporting and evaluation. We formulate a practical checklist of recommendations, delineating the characteristics that are required if a judgment prediction system is to be a valuable addition to the legal field.

1 Introduction

The task of predicting court decisions has been a topic of great interest since at least the 60s (Lawlor, 1963; Mackaay and Robillard, 1974). It has gained traction in recent years due to policies promoting publishing case law around the world, and increased efficiency and popularity of machine learning.

Today there are more than 150 academic papers claiming to predict court decisions (a.k.a. Legal Judgment Prediction) using natural language processing (NLP) and machine learning, many reporting accuracies and F1-scores of over 90% (see, for instance, Sulea et al., 2017; Varga et al., 2021; Sert et al., 2022, among others).

The media has speculated about the advent of robojudges and automated lawyers for years (Griffin, 2016; Markou, 2020; Kelly, 2022). Many in the justice sector are persuaded that machine learning technologies might usefully be employed to predict the outcomes of cases (CEPEJ, 2018). Some judges believe these technologies could even make "very minor" decisions.[1]

Yet, these systems are not widely used. While there are a number of ethical considerations (Leins et al., 2020) including issues of bias in the data (Angwin et al., 2016) that could be the reason for the hesitancy, in this paper we will demonstrate that there may be a different root cause, namely that the setup of the experiments presented in the overwhelming majority of academic papers does not allow the systems to do what they claim to be doing - *predicting court decisions*.

2 Related work

The field of Legal Judgment Prediction (LJP) has rapidly expanded over the years, with over 150 papers published in the last 10 years. The majority of the papers that we reviewed use the *text* of historical decisions to make predictions on unseen data. Some do not use NLP, but rely on manually extracted variables to make the predictions (notably, Katz et al. (2017). We looked specifically at papers in English that conduct experiments on *court* decisions, rather than any other legal decisions (excluding, for instance Branting et al. (2021) focused on WIPO decisions).

Of all the countries whose court cases feature among the reviewed papers, China has the highest proportion overall (78 papers). China publishes millions of cases online, many of which are available as part of the CAIL2018 dataset (Xiao et al., 2018) with over 2 million cases. Bench-

[1] https://www.judiciary.uk/wp-content/uploads/2023/06/Law-Society-Scotland-Law-and-Tech-Conference-2023.pdf

mark datasets attract many to develop new methods and improve the scores, with LexGLUE ECHR (Chalkidis et al., 2022) among the more prominent ones for the European Court of Human Rights (ECtHR). In fact, at least 27 papers we reviewed attempt to predict decisions of the ECtHR, with one providing access to a regularly updated website (jurisays.com) that predicts decisions in real-time.

A work that provides an even more ready-to-use tool is Galli et al. (2022). The paper includes a link to a website where one can input one's claims and receive a prediction for Italian and Bulgarian VAT cases.[2]

With so much published research, predicting court decisions (at least for some courts) may seem like a solved task, e.g. Sulea et al. (2017) achieves 99% accuracy for predicting decisions of French courts, Quemy and Wrembel (2022) achieves 98% accuracy for the judgments of the ECtHR.

3 Getting your 'facts' right

In this paper we focus on issues with the way that research in LJP is conducted today. The main issue, which concerns the experimental setup, was previously addressed in Pasquale and Cashwell (2018), (focusing on Aletras et al. (2016) paper), and later in Medvedeva et al. (2023) (reviewing a larger corpus of 27 papers). We are going to revisit this issue taking account of 171 papers in the field.

Presently, predictive systems predominantly rely on case facts extracted from judgments. Specifically, these machine learning systems are fed the text of the 'facts' section of the cases (and sometimes also the reasoning of the court) as inputs together with the corresponding decisions (e.g. violation/no violation of an article of the European Convention on Human Rights) as labels. The systems are trained to 'predict' decisions; they learn patterns within the input which correspond to particular labels. They are then evaluated on a test set (i.e. 'unseen' case facts that were not included in the training set). The 'facts' that comprise the test set, like those of the training set, are extracted from published final judgments. The labels used to evaluate the systems' performance on the test set can also be found in those judgments. While such systems might achieve high performance, they are never actually tested on 'real' data, that is data that is genuinely available to potential end-users before the judgment is reached or published. The actual user of these systems, let's say a lawyer looking to advise a client about the likely outcome of a court hearing, does not have access to the same formulation of the facts as is available and set out in the judgment. That formulation cannot exist until the judgment is made.

One might imagine that the 'facts' of the case set out in a judgment are the same as, equivalent to, or a reasonable proxy for the factual information available prior to the court's decision. However, this is incorrect. Facts very often *emerge* in the process of litigation. One side's account of the facts may (and typically will) be countered by the other side's account. Each side's account may be iteratively refined and expanded to respond to the other's. Factual evidence given by witnesses at a hearing may differ from earlier accounts given by those same witnesses. If, as is generally the case in lower courts, there is a dispute about the facts, the court will have to make a decision about the facts. In addition, courts are only concerned with those facts that are relevant for the decision in the case, taking account of the nature of the dispute or issue. That is, the courts are only concerned with those facts that relate to the legal rules which the court judges to be relevant. As a result the facts described in the judgment are often "highly selective summaries of the original case record, written by the decision-makers themselves and tailored for consistency with the decision." (Tippett et al., 2021). For all these reasons the 'facts' set out in the judgment may be very different - in substance and in form - from earlier accounts of the facts.

Some research confirms that the way in which the 'facts' of a case are formulated has a material effect on the performance of predictive systems (Tippett et al., 2021; Branting et al., 2020). Medvedeva et al. (2021) also experimented with the ECtHR data by comparing the results of training the model on the facts extracted from the final judgments (not available prior to decision-making) and training it on the facts as published by the court just after receiving the submission of the alleged victim of human right violation. Their work shows that models that perform well on the 'facts' extracted from final judgments do not perform nearly as well when trained on the 'real' data, available prior to the decisions. The best model (Hierarchical BERT) achieves an F1-score of 0.92 on final judgments and only 0.64 on facts available prior to the decision.

[2] https://adele-tool.eu

Systems trained and tested [3] on 'facts' extracted from judgments can only model that the facts described in a judgment correspond to a decision that *has already been made*; they do not in any way allow users to actually *predict* any judgments.

Although it is usual for LJP systems to rely on the 'facts' section of cases, some systems may use different information. For instance, Collenette et al. (2020) and Collenette et al. (2023) claim to predict decisions of the ECtHR by creating an Abstract Dialectical Frameworks decision tree based on factors the court considers when making decisions regarding the right to a fair trial. However, when used to 'predict' the outcome of new cases, the system depends on the user answering complex questions about 'base factors' (the leaf nodes of the decision tree) which can only be authoritatively decided by a judge. It depends, not on pre-existing information, but on interpretative decisions made by the user. The system is therefore unable to do the prediction without having part of what needs to be decided fed into it.

We limited our analysis to LJP papers published in the past 10 years (2014 until August 2023). Given the amount of papers published on the topic, our list might not contain all of the published research in the field, but we expect that we found most of the available papers. We relied on descriptions of the datasets presented in the papers in order to establish the data that the authors used for the experiments. We present the list of all reviewed papers, including information about their data and performance at `https://shorturl.at/pxKR3`. Out of 171 reviewed papers, we found only 12 ($\sim$7%) that use appropriate data for predicting decisions, with only 7 using text-based input. Additionally we found 3 papers that in principle might provide a way to predict future court decisions, but because their data is private and inaccessible, it is unclear whether all of the data used for the testing was available before the relevant judgment and therefore suitable for prediction of future outcomes. We discuss all 15 papers in section 5.

[3] In principle, one can *train* a machine learning system using various data sources, including the details within the final judgments or any information derived from them. Having additional, albeit imperfect data could potentially enhance the system's performance. However, it is imperative that *testing* is consistently performed using data accessible before the decision-making process takes place.

4 Prediction in law v. prediction in NLP

We hesitate to describe the activity of lawyers in advising about possible outcomes of cases as 'prediction'. Lawyers often qualify advice about outcomes by observing (correctly) that the law might change, that there may be room for different interpretations of the relevant law, that their client's account of the facts might not be accepted, that new issues (of fact and/or law) might arise in the course of the court proceedings. Lawyers will rarely offer a binary 'win or lose' prediction, a quantified confidence rating, or a prediction which is not hedged about with appropriate caveats (Vagts, 1978). This hesitation should not be mistaken for failure: the life of the law is its ability to adapt; the processes and procedures through which the facts of a case emerge - gradually and iteratively - are crucial for preserving the ability of the citizen to articulate their case (Waldron, 2011). Lawyers are both less and much more than fortune tellers. They have an active role to play in developing the law. Nevertheless their role in advising about the possible outcomes of cases *is* 'prediction' in the sense that they are concerned with *future* outcomes of cases. If lawyers want to know the outcome of an already decided case they can read it in the judgment.

Some few of the papers we reviewed use NLP to output a classification (e.g. violation/no violation) in response to a textual account of the facts which pre-dates the judgment. These systems are designed to predict a future outcome. However, most of the papers use NLP to output a classification in response to a textual account of facts extracted from an *existing* judgment. They classify facts in a judgment as associated with a specific verdict (i.e. a label), but the actual verdict can be found in the same judgment. There is therefore a mismatch between the usual framing of the task of LJP in NLP and a lawyer's understanding of 'prediction' of judgment. This mismatch might be understood as one of terminology. The machine learning community uses the word 'prediction' as a substitute for 'classification'; 'prediction' in this sense does not imply looking into the future. Lawyers, on the other hand, understand 'prediction' to entail prediction of some future event. For this reason Medvedeva et al. (2023) suggested researchers avoid the term 'prediction' and clearly differentiate between 'outcome identification, outcome-based judgement categorisation and outcome forecasting'. However, the mismatch is more deep-seated and

significant than one of terminology. It concerns the nature of the task performed by LJP models and the utility of the models for the intended end-user. If these are to have real-world utility for lawyers, citizens or others (e.g. political scientists) who wish to anticipate the outcome of cases they must be able to offer *prediction of future judgments*. In other words, they must offer a prediction of what the judgment *will be*, not 'predict' what the judgment *was*.

Much current research relating to LJP pays little attention to who, precisely, might use these systems and for which use cases. The majority of papers simply state that these systems could be useful to 'legal professionals' without differentiating between, for example, judges, paralegals, prosecutors, defence lawyers, clerks of court. Each of these might have different reasons for using these systems, different obligations and professional responsibilities, and therefore different use cases. The possibility that citizens might use these systems is frequently overlooked. However, as a matter of practicality, if one is to build an LJP system which has utility for some class or classes of user, it is essential to consider why they might want to use the system and what data they might use as inputs to the system. Different users may have access to different kinds of information, in different formats and languages, from different sources, reflecting different versions of events, and at different stages before or during court proceedings. An LJP system built for a specific class of end-users needs to be trained and evaluated on the data that would be available to such users. Otherwise the results do not tell one anything about how the system would perform in a real-life scenario and the system may have little or no utility.

For example, a system built for prospective applicants to the ECtHR and intended for use before an application is submitted to the court would have to be trained and tested on data available to the applicant at that stage. Such data might include textual descriptions of the facts (and arguments) formulated by the applicant, or factual accounts extracted from earlier decisions in cases raised by the applicant in domestic courts. This data has limitations; the descriptions will only contain the version of events and arguments made on behalf of the applicant and (like the factual accounts extracted from the lower courts) would normally be in the language of the applicant's country of residence. This means that the system would have to be able to handle multiple languages or be designed for a specific country. On the other hand, let's say the system is built for a big law firm which wants to take account of the outcomes predicted by the system when advising clients. Again the system would have to be trained and tested on the kind of input data available to the user at the point of use of the system for the purposes of prediction. In the case of the law firm, this might consist of the firm's written submissions for hearings, with the system being evaluated against 'unseen' written submissions and the corresponding outcomes. Like Golumbia (2015) we believe that responsibility for judging should rest with humans, not with machines. Judges, should be independent, base their decisions on law, and should not defer to machines. We are therefore very much of the view, along with Bex and Prakken (2021), that these systems should not be used for decision making by judges or indeed as a means of allowing (or compelling) a judge to assess whether their proposed decision aligns with the predictions offered by such systems. If however, the system were to be developed for such purposes, solely from an experimental setup perspective, the data would have to reflect the data available to the judge *before* they issued their judgment. In all such cases the training data, test data and finally data used for predictions would have to be obtained and used in a way that respects copyright, confidentiality, data protection and relevant rules of court. Lawyers, judges and other legal professionals would have to ensure that use of such systems is in line with their professional obligations, codes of practice and legal and ethical norms. These considerations are relevant for the design of such systems and for the choice of input data whether for training, testing or prediction. Making clear who the end-user is, even if the system is not ready to be used, rather than doing research for the sake of research, paves the way for a more appropriate experimental setup.

5 Data

Unfortunately, while the text of a judgment is not appropriate data for making predictions for that judgment, it can be hard for researchers to find *good* data for LJP. In an ideal world researchers would use the same information as is available to the court and/or the parties, e.g. statements of victims or submissions by the parties. However,

this data is not commonly available, with access to court records varying across jurisdictions. It may be restricted to physical (rather than online) access, to cases in higher courts, to certain kinds of case documents, or provided only for a fee (Naglič et al., 2013).

The papers that use appropriate data for future predictions consistently exhibit much poorer performance compared to those relying on flawed data. Among papers employing text-based input for training, only one achieves over 70% accuracy. Specifically, Medvedeva et al. (2020) attain 75% accuracy in predicting ECtHR decisions based on summaries from so-called communicated cases, prepared by the court, derived from applicant submissions, and published before the judgments are made. However, a study with a similar albeit smaller dataset subsequently reports a reduced accuracy rate of 65%-68% depending on the year of the test data (Medvedeva et al., 2021).

Note that these summaries are only available if and when the ECtHR accepts an application and only for a subset of cases. Consequently, depending on the end-user and their specific requirements, these summaries may still not suffice for making predictions. Since applicants do not have the case summary before they submit the applications to the court, it is suitable for prediction only once the legal proceedings begin. Then applicants may use the case summary as input to an LJP system to estimate their prospects of success, potentially opting for a settlement to avoid lengthy proceedings.

Using lower court decisions to predict higher court outcomes as explored by Waltl et al. (2017) and Jacob de Menezes-Neto and Clementino (2022) offers a way to use pre-judgment data, yet performance has been modest with the papers reporting F1-score of 57% and a Matthews Correlation Coefficient (MCC) of 0.37, respectively. Predicting court decisions is an inherently challenging task. The challenge is compounded by the fact that many countries and courts only publish a subset of their cases at each level, making it difficult to access sufficient data from both lower and higher courts.[4] It is crucial to acknowledge, however, that not all cases eligible for appeal to a higher court will actually pursue this course of action, introducing potential selection bias when training and evaluating the system.

Other papers that use text as input rely on various documents from different stages of legal proceedings. For example, Semo et al. (2022) uses plaintiff claims, McConnell et al. (2021) complaint documents, and Tippett et al. (2021) legal briefs. While these documents may be available before the decision is made, they typically represent the claims of one party, and therefore may not include all relevant information. The papers report accuracy scores of 67%, 61% and an MCC of 0.48, respectively.

The papers predicting decisions of the US Supreme Court (e.g. Sharma et al., 2015; Katz et al., 2017) often use a manually annotated SCDB dataset (Spaeth et al., 2014), containing 240 expert-annotated variables, rather than text from legal documents. Such data is rarely available for other courts, and would require significant manual effort to compile. These papers do not use textual input, yet both achieve 70% accuracy. Such data could potentially be supplemented with, for example, decisions of the lower court and/or expert opinions. For instance, Kaufman et al. (2017) use oral argument in combination with the annotated variables, therefore creating a system that can predict decisions once the court case started, reporting 74% accuracy. Chen and Eagel (2017) incorporate seemingly unrelated data like weather and news trends to improve the performance (up to 79% accuracy).

Similarly, published judgments, although unsuitable for directly predicting judgments, could serve as a source from which to extract factors or variables that are relevant to the outcomes of those cases. These factors, in turn, might be used for prediction. For instance, Hsieh et al. (2021) extract information about plaintiff and defendant, including their income and debt information, to make predictions about discretionary damages. However, it is not always obvious whether the factors are available in advance or will have to be established by the court. For example, França et al. (2020); dos Santos et al. (2020); Bagherian-Marandi et al. (2021) who also adopt this approach mention using client information, as well as extracting details from judgments. However, since their data is not published or available for examination the precise nature of the extracted features remains unclear. Bagherian-Marandi et al. (2021), for instance, has

[4] Some countries e.g. Brazil reportedly publish the majority of their case law (Jacob de Menezes-Neto and Clementino, 2022). Similarly, Danish Supreme Court cases are published together with underlying decisions of the lower courts, while the US Supreme Court publishes most case filings including parties' briefs.

the value of the claim as one of the features. This feature could relate either to the sum claimed in the initial submission (appropriate data) or the sum determined during the course of the proceedings, with the latter not being available prior to the judgment (inappropriate data). Therefore, the performance of these systems should be tested on data provided by the users rather than relying on extracted data. This approach ensures that the system's performance mirrors a real-world scenario.

Given the scarcity of papers conducting research on appropriate data and the relatively modest performance of many of those systems, it is evident that there is substantial room for investigation to establish the most effective approaches for LJP. We have offered examples of data that might be appropriate for LJP ('good' data). However further research is needed to establish what 'good' data is available to researchers in different jurisdictions, and, of this data, which may produce optimal results. Crucially however, and contrary to the approach adopted in the vast majority of papers, doing LJP 'right' depends on selecting and using data that predates the judgment.

6 Benchmark datasets

As we have discussed, the facts of the judgments that one is predicting are the wrong data for the LJP task. It follows that benchmark datasets, such as CAIL2018 (Xiao et al., 2018), LexGLUE ECHR A & B (Chalkidis et al., 2022) or ILDC CJPE (Malik et al., 2021) whose input data consists of facts extracted from judgments, are of little or no assistance for the task of prediction of future judgments. Models which predict a future judgment can only be benchmarked on a dataset containing textual representations of facts that are available before the judgment.

Endorsement by the NLP community of datasets such as CAIL2018, LexGLUE ECHR and ILDC CJPE as benchmarks for prediction of judgment has ramifications. Koch et al. (2021) note that "When they institutionalize benchmark datasets, task communities implicitly endorse these data as meaningful abstractions of a task or problem domain. The institutionalization of benchmarks influences the behavior of both researchers and end-users." CAIL2018, LexGLUE ECHR, ILDC CJPE and similar datasets implicitly signal to researchers, lawyers and policy makers that it is meaningful to carry out the task of prediction of judgment by using data extracted from cases in which the judgment is already available as the model input. They also implicitly signal that the performance scores of models tested on the datasets are meaningful indicators of performance in the real-world task of prediction of (future) judgment.

Do the 'predictions' output by models tested against such benchmarks have any real-world utility? We suggest they do not. Such 'predictions' involve labelling or categorising input data as correlated with a particular (already known) outcome (Medvedeva et al., 2023). These 'predictions' have no informational value for lawyers, their clients or others who wish to obtain a prediction of a future judgment (Bex and Prakken, 2021; Medvedeva et al., 2023). Moreover, to our knowledge there is no research indicating that any improvements of the results on such a dataset would translate to improved performance in a real-world context. In fact, we are only aware of research that has demonstrated the contrary (see Medvedeva et al. (2021).

To mitigate these issues, we propose that the ideal test set should - at a minimum - consist of data in which no information has been derived from the judgments that one is predicting. Instead, this data should be gathered from the parties involved in the case or from documents accessible before the decision-making process. Conducting an experiment in which data is gathered for cases that are still pending decisions, followed by a waiting period until the court issues its judgments for comparison, would be the most dependable method to safeguard against flawed test results. Although this approach may present increased difficulties due to its longitudinal nature, it would guarantee the reliability of the test results.

7 Addressing explainability

The real-world utility of LJP systems may also depend on their explainability. "[I]n Law the explanation is what matters" (Bench-Capon, 2021). When lawyers give advice about outcomes, they generally explain why an outcome or range of outcomes might be likely. They explain how the law may apply to the client's problem. Similarly, judges usually offer explanations which serve to justify their decisions with reference to the facts found to be established and the relevant law. This aspect of judicial practice relates to the public character of law and the ability of citizens to engage with the law (Waldron, 2008). Both lawyers and judges

provide explanations which involve the exercise of legal reasoning (Branting, 2020). These explanations have a particular form. They link the facts of the case and the law to the conclusion; they are grounded in law. We also expect these explanations to meet a qualitative standard; we want them to accurately represent the law and to demonstrate sound reasoning.

Is it necessary that LJP systems issue explanations that are grounded in law? This may depend on how, by whom and for whose benefit these systems will be used. Despite our reservations, these systems might be used in place of judges, to make decisions rather than merely offering predictions. In that scenario, it would seem essential for such a system to be capable of providing an explanation of its output that is grounded in law (Završnik, 2020). Otherwise, how could one assess whether the output was justified according to legal standards, or how, and on what legal basis, litigants might challenge the output of the system? Similarly, if persons without legal representation use these systems to estimate their prospects of success in a court case, it would seem essential for the system to provide an explanation grounded in law. How else could such persons make sense of those outputs? If on the other hand lawyers use these systems as additional tools, it might be desirable but not essential for the system to provide explanations grounded in law. Lawyers should provide their own explanations of the outcome or outcomes they consider likely.

A range of approaches to explainability can be found in the LJP literature. For example, machine learning systems which frame LJP as a classification task may output an 'explanation' about which features the system treats as important. The system may output descriptions or visualisations of which words, sentences or paragraphs of the input data (usually the 'facts' of a case) contributed to the output predictions (Medvedeva et al., 2020; Malik et al., 2021). Alternatively, machine learning systems that treat LJP as a text generation task may generate text in the form of steps of reasoning (a chain of thought) or a legal syllogism together with a conclusion (a prediction) (Jiang and Yang, 2023). The 'reasoning' steps serve as the explanation. Hybrid systems which combine machine learning with symbolic reasoning approaches may output justifications of the predictions output by the system (Prakken and Ratsma, 2022). These justifications would not represent the 'reasoning steps' of the predictive model though they may provide an ex post explanation of the prediction.

The focus on explainability is welcome, and certain of these approaches may hold some promise. However, for most users and use cases, the explanations (if they are to be useful) must be grounded in law, relate to predictions of future judgments and evaluated for soundness by suitable methods. We are not aware of any LJP system that outputs 'explanations' that meet all three criteria. This explainability deficit is a significant limitation which is relevant for the evaluation of LJP models, for their real-world utility and for the potential benefits, harms and impacts associated with their use.

8 An ethical approach to model evaluation and reporting

We have suggested that current practice in relation to the evaluation of LJP models is frequently flawed, involving a race-to-the-top against benchmark datasets unsuited for the task of prediction of future judgments and a failure to consider intended users and use cases. If these issues are relevant for training and testing a system, they are also relevant for further evaluation and reporting. We favour an approach to evaluation which is "application-centric" (Hutchinson et al., 2022), that is, an approach which considers the fitness of the model for its real-world application context. An application-centric approach would necessarily entail a focus on the intended user and use case and highlight the unsuitability of LJP benchmark datasets comprising formulations of 'facts' only available at the point of judgment.

An application-centric approach takes account of direct benefits and harms to users and those affected by the outputs of the system (Hutchinson et al., 2022). Adoption of this approach to evaluation therefore requires a good understanding of the context in which a model is likely to be used, who will benefit or suffer adverse consequences as a result of such use and how those consequences are likely to arise. It requires a sufficient understanding of the way in which LJP systems might shape lawyers' advice and judges' determinations, ultimately affecting legal outcomes for citizens, and a recognition that these systems are "intended to inform decisions about matters central to human life or flourishing" (Mitchell et al., 2019). More concretely, it requires an appreciation that a citi-

zen might suffer harm if, for example, a system suggests a claim will succeed when the claim is doomed to failure on legal grounds, or provides a judge with an overly harsh recommendation in relation to sentencing.

Crucially an application-centric approach takes into account the potential impact of "changes to the ecosystem itself" (Hutchinson et al., 2022). In the case of LJP systems this entails reflection on how the use of these systems may affect the ecosystems of law and legal practice. For example, if judges were to align their decisions with the outputs of LJP systems, this might lend credibility to the notion that machines can engage in legal reasoning, reduce the human element in decision-making, devalue legal reasoning and de-skill judges. Lawyers who use these systems might be "encourage[d] ... to base their litigation strategy on factors other than the legal merits of the case" (Diver et al., 2022). Law might stagnate; as Bench-Capon (2021) notes, the model producing the prediction can only be "trained on past decisions". There is a risk of confusing or conflating what LJP systems can do with the exercise of legal reasoning within a legal institutional framework which allows for "interpretation, contestation and argumentation" (Hildebrandt, 2019). The impacts are likely to be amplified if the systems are used for decision-making rather than decision-support. In that context, the inability of LJP systems to provide explanations grounded in law could have serious ramifications. Of course, the use of these systems might also produce benefits - the point is to reflect on the impact of the research and the technology on the application ecosystem.

We suggest that an application-centric approach is both necessary and appropriate where researchers maintain that the systems they develop have real-world utility. It encourages trust; instead of merely claiming real-world utility, researchers would demonstrate their attention to the context of application. It is "better aligned with the needs of decision-makers who consider whether to use a model in an application." (Hutchinson et al., 2022). Model cards, (Mitchell et al., 2019) which allow for reporting on the intended user, intended use case, the benefits, harms and social impacts, might be of use in this context. We encourage the use of such a reporting framework as a means of clearly and transparently communicating evaluations of models which look beyond accuracy metrics.

9 Discussion

Despite the perception of prolific research activity, the reality is that the LJP field has produced only a meagre 12 papers to date, covering 9 courts, and reporting performance scores between 56% and 79% accuracy. By diverting resources and attention towards systems that do not deliver on their promises, researchers and developers may be discouraged from pursuing more appropriate and impactful avenues. The atmosphere of competition, with numerous models claiming accuracy rates exceeding 90%, may hinder the publication of rigorous initial-stage research and incremental advancements in the field that produce much lower performance. Consequently, it impedes the overall progress of LJP research.

Such faulty research not only hampers scientific progress but also poses potential risks to individuals. For instance, a system such as ADELE, while conceived as an academic project, can be used to make predictions based on claims provided by users, even though it was not trained or evaluated on such data. Given the inflated scores of systems claiming to be predicting court decisions, there is a distinct possibility that these systems may eventually find their way onto the market and into courts.

We propose a number of recommendations to address the concerns outlined in this paper. We recommend that researchers building and reporting on systems that claim to do LJP identify the end-user, the application of the system and the stage in the legal process when the system would be used. Researchers should then evaluate whether the data they plan to use is available to the user at that stage of the proceedings or collect a new dataset. Depending on the purpose of the application one may choose to create a system that offers some form of explainability, whether in the sense of explaining how the model makes the prediction (for instance, for error analysis and correction) or, explanations that take the form of legal reasoning. When building such a system, even within academic research, one should keep in mind and report not only potential benefits, but also potential harms that might result from use of the system. We encourage developers to also consider the broader impact that the system may have on individuals, law, legal procedures and society as a whole. We summarise these recommendations in a schematic checklist available in Appendix A.

10 Conclusion

Our paper delves into the current landscape of Legal Judgment Prediction (LJP) research and its issues. Through examination of 171 papers, we find that only a small fraction of LJP papers which claim to predict court decisions achieve this objective. The majority falter by using the wrong data for the task at hand. Despite many papers reporting exceptionally high performance, it becomes evident that the task of LJP proves to be considerably more challenging when researchers use an appropriate experimental setup.

In addition, we scrutinise when and for whom LJP has real-world utility. We emphasise the importance of considering the end-user and the use case when developing the models. The prevalent reliance on readily available data and benchmark datasets is not in line with this focus. We advocate for a shift in approach to redirect the field away from a potentially futile trajectory, ultimately preventing the misallocation of resources on endeavors that lack practical utility within the legal domain.

Acknowledgements

The second author, Pauline McBride, is funded by the European Research Council (ERC) under the HORIZON2020 Excellence of Science programme ERC-2017-ADG No 788734

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. *PeerJ Computer Science*, 2.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: there's software used across the country to predict future criminals. and it's biased against blacks. ProPublica 2016. Https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Navid Bagherian-Marandi, Mehdi Ravanshadnia, and Mohammad-R Akbarzadeh-T. 2021. Two-layered fuzzy logic-based model for predicting court decisions in construction contract disputes. *Artificial intelligence and law*, pages 1–32.

Trevor Bench-Capon. 2021. The need for good old fashioned AI and law. *Jusletter-IT*, (fses):23–35.

Floris Bex and Henry Prakken. 2021. On the relevance of algorithmic decision predictors for judicial decision making. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 175–179.

Karl Branting, Carlos Balhana, Craig Pfeifer, John S Aberdeen, and Bradford Brown. 2020. Judges are from Mars, pro se litigants are from Venus: Predicting decisions from lay text. In *JURIX*, pages 215–218.

L Karl Branting. 2020. Explanation in hybrid, two-stage models of legal prediction. In *The 3rd XAILA Workshop at JURIX*.

L Karl Branting, Craig Pfeifer, Bradford Brown, Lisa Ferro, John Aberdeen, Brandy Weiss, Mark Pfaff, and Bill Liao. 2021. Scalable and explainable legal prediction. *Artificial Intelligence and Law*, 29:213–238.

European Commission for the Efficiency of Justice CEPEJ. 2018. European ethical charter on the use of artificial intelligence in judicial systems and their environment. *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment*.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Daniel L Chen and Jess Eagel. 2017. Can machine learning help predict the outcome of asylum adjudications? In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 237–240.

Joe Collenette, Katie Atkinson, and Trevor Bench-Capon. 2020. An explainable approach to deducing outcomes in European Court of Human Rights cases using ADFs. *Frontiers in Artificial Intelligence and Applications*, 326:21–32.

Joe Collenette, Katie Atkinson, and Trevor Bench-Capon. 2023. Explainable AI tools for legal reasoning about cases: A study on the European Court of Human Rights. *Artificial Intelligence*, 317:103861.

Laurence Diver, Pauline McBride, Masha Medvedeva, Arjun Bhubaneshwar Banerjee, Eva D'hondt, Tatiana Duarte Nicolau, Desara Dushi, Gianmarco Gori, Emilie Van Den Hoven, Paulus Meessen, et al. 2022. Typology of legal technologies. In *Cross-disciplinary Research in Computational Law (CRCL): Computational'Law'on Edge*.

Pedro TC dos Santos, Fernando Henrique, Venicius Garcia, Victor RS Ferreira, Antonino C dos Santos Neto, Johnatan C Souza, Caio Manfredini, João VF França, José MC Boaro, Geraldo Braz Junior, et al. 2020. Multiclass legal judgment outcome prediction for consumer lawsuits using XGBoost and TPE. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 881–886. IEEE.

João VF França, José MC Boaro, Pedro TC dos Santos, Fernando Henrique, Venicius Garcia, Caio Manfredini, Domingos AD Júnior, Francisco YC de Oliveira, Carlos EP Castro, Geraldo Braz Junior, et al. 2020. Legal judgment prediction in the context of energy market using gradient boosting. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 875–880. IEEE.

Federico Galli, Giulia Grundler, Alessia Fidelangeli, Andrea Galassi, Francesca Lagioia, Elena Palmieri, Federico Ruggeri, Giovanni Sartor, and Paolo Torroni. 2022. Predicting outcomes of Italian VAT decisions. In *Legal Knowledge and Information Systems*, pages 188–193. IOS Press.

David Golumbia. 2015. Judging like a machine. In *Postdigital aesthetics: art, computation and design*, pages 123–135. Springer.

Andrew Griffin. 2016. Robot judges could soon be helping out with court cases. Accessed: 11-08-2023.

Mireille Hildebrandt. 2019. Data-driven prediction of judgment. Law's new mode of existence?

Decheng Hsieh, Lieuhen Chen, and Taiping Sun. 2021. Legal judgment prediction based on machine learning: Predicting the discretionary damages of mental suffering in fatal car accident cases. *Applied Sciences*, 11(21):10361.

Ben Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. 2022. Evaluation gaps in machine learning practice. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1859–1876.

Elias Jacob de Menezes-Neto and Marco Bruno Miranda Clementino. 2022. Using deep learning to predict outcomes of legal appeals better than human experts: A study with data from Brazilian federal courts. *PloS one*, 17(7):e0272287.

Cong Jiang and Xiaolei Yang. 2023. Legal syllogism prompting: Teaching large language models for legal judgment prediction. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 417–421.

Daniel Martin Katz, Michael J Bommarito II, and Josh Blackman. 2017. A general approach for predicting the behavior of the Supreme Court of the United States. *PloS one*, 12(4).

Aaron Kaufman, Peter Kraft, and Maya Sen. 2017. Machine learning, text data, and Supreme Court forecasting. *Project Report, Harvard University*.

Jemima Kelly. 2022. AI-driven justice may be better than none at all. Accessed: 11-08-2023.

Bernard Koch, Emily Denton, Alex Hanna, and Jacob G Foster. 2021. Reduced, reused and recycled: The life of a dataset in machine learning research. *arXiv preprint arXiv:2112.01716*.

Reed C Lawlor. 1963. What computers can do: Analysis and prediction of judicial decisions. *American Bar Association Journal*, pages 337–344.

Kobi Leins, Jey Han Lau, and Timothy Baldwin. 2020. Give me convenience and give her death: Who should decide what uses of NLP are appropriate, and on what basis? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2908–2913, Online. Association for Computational Linguistics.

Ejan Mackaay and Pierre Robillard. 1974. *Predicting judicial decisions: The nearest neighbour rule and visual representation of case patterns*, volume 3(3-4).

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripa Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. *arXiv preprint arXiv:2105.13562*.

Christopher Markou. 2020. Are we ready for robot judges? Accessed: 11-08-2023.

Devin J McConnell, James Zhu, Sachin Pandya, and Derek Aguiar. 2021. Case-level prediction of motion outcomes in civil litigation. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 99–108.

Masha Medvedeva, Ahmet Üstun, Xiao Xu, Michel Vols, and Martijn Wieling. 2021. Automatic judgement forecasting for pending applications of the European Court of Human Rights. In *Proceedings of the Fifth Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2021)*.

Masha Medvedeva, Martijn Wieling, and Michel Vols. 2023. Rethinking the field of automatic prediction of court decisions. *Artificial Intelligence and Law*, 31(1):195–212.

Masha Medvedeva, Xiao Xu, Martijn Wieling, and Michel Vols. 2020. JURI SAYS: Prediction system for the European Court of Human Rights. In *Legal Knowledge and Information Systems: JURIX 2020: The Thirty-third Annual Conference, Brno, Czech Republic, December 9-11, 2020*, volume 334, page 277. IOS Press.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Vesna Naglič et al. 2013. National practices with regard to the accessibility of court documents.

Frank Pasquale and Glyn Cashwell. 2018. Prediction, persuasion, and the jurisprudence of behaviourism. *University of Toronto Law Journal*, 68(supplement 1):63–81.

Henry Prakken and Rosa Ratsma. 2022. A top-level model of case-based argumentation for explanation: formalisation and experiments. *Argument & Computation*, 13(2):159–194.

Alexandre Quemy and Robert Wrembel. 2022. ECHR-OD: On building an integrated open repository of legal documents for machine learning applications. *Information Systems*, 106:101822.

Gil Semo, Dor Bernsohn, Ben Hagag, Gila Hayat, and Joel Niklaus. 2022. ClassActionPrediction: A challenging benchmark for legal judgment prediction of class action cases in the US. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 31–46.

Mehmet Fatih Sert, Engin Yıldırım, and İrfan Haşlak. 2022. Using artificial intelligence to predict decisions of the turkish constitutional court. *Social Science Computer Review*, 40(6):1416–1435.

Ranti Dev Sharma, Sudhanshu Mittal, Samarth Tripathi, and Shrinivas Acharya. 2015. Using modern neural networks to predict the decisions of Supreme Court of the United States with state-of-the-art accuracy. In *International Conference on Neural Information Processing*, pages 475–483. Springer.

Harold Spaeth, Lee Epstein, Ted Ruger, Keith Whittington, Jeffrey Segal, and Andrew D Martin. 2014. Supreme Court database code book.

Octavia-Maria Sulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P Dinu, and Josef Van Genabith. 2017. Exploring the use of text classification in the legal domain. In *In Proceedings of the 2nd Workshop on Automated Semantic Analysis of Information in Legal Texts (ASAIL 2017)*.

Elizabeth C Tippett, Charlotte S Alexander, Karl Branting, Paul Morawski, Carlos Balhana, Craig Pfeifer, and Sam Bayer. 2021. Does lawyering matter? Predicting judicial decisions from legal briefs, and what that means for access to justice. *Tex. L. Rev.*, 100:1157.

Detlev F Vagts. 1978. Legal opinions in quantitative terms: The lawyer as haruspex or bookie. *Bus. Law.*, 34:421.

Dávid Varga, Zoltán Szoplák, Stanislav Krajci, Pavol Sokol, and Peter Gurskỳ. 2021. Analysis and prediction of legal judgements in the Slovak criminal.

Jeremy Waldron. 2008. The concept and the rule of law. *Ga L Rev*, 43.

Jeremy Waldron. 2011. The rule of law and the importance of procedure. *Getting to the Rule of Law*, 3:4–5.

Bernhard Waltl, Georg Bonczek, Elena Scepankova, Jörg Landthaler, and Florian Matthes. 2017. Predicting the outcome of appeal decisions in Germany's tax law. In *International Conference on Electronic Participation*, pages 89–99. Springer.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. CAIL2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Aleš Završnik. 2020. Criminal justice, artificial intelligence systems, and human rights. In *ERA forum*, volume 20, pages 567–583. Springer.

A Appendix A: Checklist for developing Legal Judgment Prediction systems

End-User:

☐ Clearly establish the end-user of the system.

System Application:

☐ Define the specific application and purpose of the system within the legal process.

☐ Determine the stage within the legal process where the system would be used (e.g. prior to going to court, during proceedings).

Data Evaluation:

☐ Assess whether the test set mirrors the data available to the user at the relevant stage of the legal proceeding according to the defined application.

Explainability:

☐ Evaluate whether the system needs to provide explanations for its predictions given the application and, if so, specify the nature of the explanations required.

Ethical considerations:

☐ Identify potential benefits or harms related to system performance, such as (incorrect) predictions affecting legal proceedings, individual's lives and legal protections.

☐ Consider potential broader repercussions of system design choices, including impacts on individuals, law, legal procedures, and society at large.