Taking the Law More Seriously by Investigating Design Choices in Machine Learning Prediction Research

Cor Steging[1,*], Silja Renooij[2] and Bart Verheij[1]

[1]Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence, University of Groningen
[2]Department of Information and Computing Sciences, Utrecht University

Abstract

Approaches to court case prediction using machine learning differ widely with varying levels of success and legal reasonableness. In part this is due to some aspects of law, such as justification, being inherently difficult for machine learning approaches. Another aspect is the effect of design choices and the extent to which these are legally reasonable, which has not yet been extensively studied. We create four machine learning models tasked with predicting cases from the European Court of Human Rights and we perform experiments in order to measure the role of the following four design choices and effects: the choice of performance metric; the effect of including different parts of the legal case; the effect of a more or less specialized legal focus; and the temporal effects of the available past legal decisions. Through this research, we aim to study design decisions and their limitations and how they affect the performance of machine learning models.

Keywords

Court case prediction, design choices, machine learning

1. Introduction

Recently, much work has been done in the field of court case predictions. While automatically determining the outcome of court cases remains an academic exercise, the large variation in the ways that previous research has tackled the problem makes it nearly impossible to compare the approaches [1]. The law has unique characteristics, making it difficult to apply machine learning in the legal domain: machine learning is retrospective, assumes normally distributed, homogeneous data that is largely free of errors, and it often cannot explain its decision-making [2]. The law on the other hand is prospective, changes over time, contains wrong decisions, and demands arguments for the decisions made. These unique characteristics of the law are not always taken into account. To take the law more seriously, we must consider these when doing machine learning research in the field of AI & Law.

Some requirements of the law, such as justification, are inherently difficult for machine learning systems, and machine learning systems have been shown to use unsound reasoning [3]. However, despite their importance, our focus in this paper will not be on justification, responsibility or explainibility. Moreover, our goal is not to create a machine learning system that obtains a better performance, or has a better alignment with legal experts [4]. Instead, we investigate the effect of specific design choices and effects in machine learning research, in order to better analyze performance and alignment with characteristics of the legal domain.

We focus on research involving cases from the European Court of Human Rights (ECHR), which has been used as a benchmark in a number of studies. ECHR data is included in the LexGLUE benchmark datasets [5], and forms the basis of the ECHR-OD repository [6]. Previous studies have applied different machine learning systems to this dataset, using various methods and achieving different levels of success [7, 8, 9, 10, 11]. To study the effects of design choices, we train four different types of machine learning models on cases from the ECHR: an SVM, a Naive Bayes (NB) Classifier, a Random Forest (RF) and a BERT model. For these four models, we study the choice of performance metrics; the effect of including different parts of the legal case; the effect of a more or less specialized legal focus; and the temporal effects of the available past legal decisions.

Our first set of experiments focuses on the replication and expansion of results in the literature. We train and test our four models on two different datasets from the ECHR, using various parts of each case as input, and report both the accuracies and Matthew's Correlation Coefficient (MCC) on each task, model and dataset.

The ECHR covers a number of separate articles. Earlier work on court case prediction used either single, general models trained on all articles [8, 10, 5], or a separate, specialized classifier for each article [7, 9]. In the second set of experiments, we create both a Generalist model and an *Ensemble* of specialized models in order to investigate the differences in their performances.

Proceedings of the Sixth Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2023), June 23, 2023, Braga, Portugal.
*Corresponding author.
c.c.steging@rug.nl (C. Steging); s.renooij@uu.nl (S. Renooij); bart.verheij@rug.nl (B. Verheij)
0000-0001-6887-1687 (C. Steging); 0000-0003-4339-8146 (S. Renooij); 0000-0001-8927-8751 (B. Verheij)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3441/paper6.pdf

The third and final set of experiments studies temporal effects. We investigate the effects of training models on cases from the past to predict future cases, compared to models trained on randomly split data. Furthermore, we explore the effects of training on cases from varying time windows for a model that predicts future cases.

In Section 2, we discuss relevant background information. Section 3 addresses our experimental setup and Section 4 the experiments themselves. We conclude our paper in Section 5.

2. Background

The current work focuses on the cases made publicly available by the ECHR, which is an international court that deals with cases claiming violations of articles laid out by the European Convention on Human Rights. A case can pertain to multiple articles of the ECHR and multiple articles can be violated. Each case description can be divided into the following main parts:

Introduction: general information, such as title, date and details about the section of the Court.

Procedure: the course of action taken from lodging and application until the final Court judgement.

Facts: the circumstances of the case, such as the relevant background information of the applicant and other events and circumstances; and the relevant law from documents other than the ECHR.

Law: the legal arguments of the Court.

Judgement: the Court's decision.

Dissenting/Concurring opinions: judges' opinions and why they voted for or against a violation.

In court case prediction, the case text acts as the features and the judgement as the label. Three variations of the prediction task have been studied:

- In the **Binary classification task (BC)**, there is one dataset that contains all cases. Models are tasked with predicting whether *any* article was violated for each case [7, 9, 8, 10].
- In the **Multi-label classification task (MLC)**, there is one dataset that contains all cases. Models are tasked with predicting *which* articles were violated for each case [8, 10].
- In the **Article classification task (AC)**, there are multiple datasets, one for each article. Models are tasked with predicting whether a specific article was violated for each case [7, 9].

The first models applied to the ECHR classification task were Support Vector Machines (SVM) [7, 9]. A later study used BERT, a state-of-the-art pre-trained transformer model. While transformers tend to outperform traditional models, BERT yielded a lower accuracy on the ECHR task [8], because the ECHR cases greatly exceed BERT's 512 token limit and had to be truncated. Chalkidis et al. therefore also introduced an hierarchical version of BERT (HIER-BERT), where the words of each fact in the case are first converted to a *fact embedding* using the base BERT model. This version performed significantly better on the binary classification task than their regular BERT model with truncation (F1-scores of 82.0% vs. 17.0%). By pre-training this BERT model on additional legal data, a legal-BERT was developed, specifically suited to legal texts [10] (see also [12]), which performed better on the ECHR task than the HIER-BERT model (F1-scores of 88.3.0% vs. 82.0%). It has been noted, however, that specialized transformers in the legal domain (legal-BERT) provide relatively little improvement over a standard transformer, especially when compared to the difference between regular and specialized BERT models in other fields, such as in the biomedical domain[13]. Mumford et al. took a hybrid approach to the court case prediction task, opting to combine HIER-BERT models with Abstract Dialectical Frameworks. While it is difficult to compare the performance of this hybrid model to other research, it did outperform a HIER-BERT model trained on the same subset of ECHR data. Additionally, the hybrid model is more explainable and can provide justifications for its predictions.

3. Experimental setup

Here we describe our datasets, machine learning models, preprocessing steps, and performance metrics used. All of the code used to run the experiments can be found in a public repository[1].

3.1. Datasets

We train machine learning models on cases from the ECHR and use these models to predict new case decisions. We use the dataset from the ECHR Open Data project (ECHR-OD)[2] [6]. This repository contains formatted and standardized data from the ECHR that is automatically updated every month, establishing a public shared baseline for machine learning models. Each case in this dataset contains the text of the case and the outcome, i.e. which articles were considered violated. A single case can violate multiple articles.

We set up our datasets for the article classification (AC) task and the binary classification (BC) task. For the AC task, there are 9 datasets, one for each article.

[1] https://github.com/CorSteging/InvestigatingDesignChoices
[2] https://echr-opendata.eu/. Accessed 21 Nov. 2022

Table 1
Number of cases per article in the ECHR-OD dataset and percentage of violation cases per article.

Article	Size	Violation %
2	1212	80.94
3	3489	81.00
5	3165	84.27
6	8272	87.51
8	1841	71.75
10	789	76.05
11	340	84.71
13	2309	91.51
14	578	47.23
All	14910	81.55

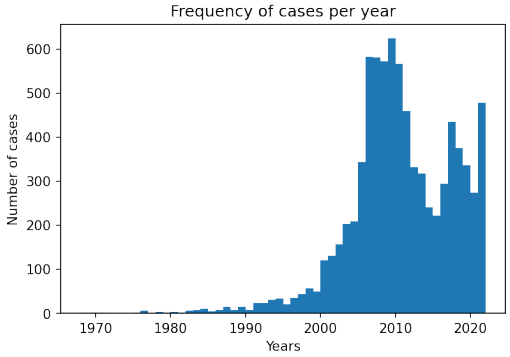

Figure 1: Frequency of cases per year in ECHR-OD dataset.

Each dataset contains all of the cases pertaining to that specific article, and the binary label indicates whether that specific article was violated. For the BC task, there is one dataset that contains all cases. The binary label of these cases indicates whether *any* article was violated. The number of cases in each dataset can be found in Table 1, alongside the percentage of cases that evaluate to a violation of their respective article. Note again that multiple articles can be considered for a single case. The sum of all datasets for each individual article in Table 1 is therefore greater than the number of cases in the 'All' dataset. The outcome in most cases is a violation. The label distribution is therefore skewed towards violation.

To train a model, we balance the dataset used such that half of the cases evaluate to violation and the other half evaluate to non-violation. To balance a dataset, we randomly remove violation cases from the dataset until their number equals the number of non-violation cases.

The version of the dataset that we use from the ECHR-OD contains 14910 cases from 1968 up to and including 2022. The distribution of the cases across the years is skewed heavily towards the more recent years, however. This is clearly visible in Figure 1, where we plot the number of cases per year. Since some of the earlier years do not contain any cases, we only include cases from 1978 until 2022 in our experiments.

3.2. Models

In our study we use four different types of models: an SVM, a Naive Bayes (NB) classifier, a Random Forest (RF) classifier and a BERT model. These are all commonly used models known for their effectiveness in text classification tasks [14]. For the SVM, we use the exact same parameters as reported in [9]. The parameters of the other models are tuned using a grid search for each experiment, where we validate the performance on an unseen part of the training set. For the SVM, NB and RF models, we use the scikit-learn library [15]. We use the BERT transformer from the open-source Hugging Face library [16] and limit the number of tokens to 512 using the default Tokenizer from that library.

3.3. Preprocessing

We train and test our four models on the ECHR-OD dataset of case texts, preprocessed to remove unnecessary information from the text and to reduce the token size for the BERT model, by applying the following heuristics:

- Change all characters to lowercase
- Remove all punctuation except for '?'
- Remove special characters, newlines and trailing white spaces
- Change 't to not ("don't" becomes "do not")
- Change all fact numbers to '>'
- Remove stop words using nltk [17]
- Remove unnecessary words that occur in every case (such as subheadings)

For the SVM, NB and RF models, the texts are then converted to n-grams and normalized using TF-IDF. The parameters for preprocessing and TF-IDF are fine-tuned using a grid search.

3.4. Performance Metrics

Most models in the literature report the classification accuracy or F1-scores of their model. While this was common practice in machine learning, more recent studies have steered away from using accuracy in favor of the Matthew's correlation coefficient (MCC) [18], which ranges from -1 (worst) to 1 (best). Contemporary measures like accuracy or even F1-scores have been shown to yield inflated results on binary classification tasks [19], especially on imbalanced datasets such as the ECHR cases. The MCC, on the other hand, is only high if all four confusion matrix categories are accurate: high true positives

and negatives, and low false positives and negatives. For example, a model that always predicts 'violation' will score an accuracy of 81.55% on the entire ECHR dataset, since 81.55% of the cases have a violation label. From a legal point of view, this is an extremely irresponsible and poorly designed model. However, the accuracy is high and the F1-score of this model would even be 89.83%, beating the state of the art. The MCC of such a model, however, will be 0, indicating that its predictive power is equal to random guessing. While a macro-averaged F1-score can be used for unbalanced data, it is known to be biased and does not take true negative predictions into account [19]. To take the law more seriously, we therefore choose to use the MCC to evaluate our models, even though we work with balanced datasets. We in general advocate the use of the MCC for binary classification as a best practice. We report the accuracy of our models when comparing their results with the results from the literature. In the rest of our study, we will report MCC values only.

4. Experiments

We now discuss our three sets of ECHR-OD experiments.

4.1. Experiment 1: Extended replication

To evaluate the performance of each model, we apply a 10-fold cross validation to the model for each article (the AC task), and for all articles at once (the BC task). We balance each of the datasets such that exactly half of the cases evaluate to 'violation'. We compare the performance on the ECHR-OD dataset to that of models from the literature. For comparison, we also train and test our models on the subset of data used by Medvedeva et al. (2020; see [9] for details), which contains only 3133 cases from 1968 to 2017.

Performance Results Table 2 shows four sets of rows containing performances of models on the AC task and the BC task (All). In the first set of rows, we show results from previous research. Note that results marked with an asterisk ($*$) are F1-scores rather than accuracies. In the second set of rows, we show the accuracies of the 10-fold cross validation of our SVM, NB, RF and BERT models using the data from [9]. The third set of rows lists the accuracies of our four models using the ECHR-OD data, and the last set of rows contains the MCC values of our four models on the ECHR-OD dataset. Note that we decided to round off our results values to one decimal, just as in the more recent studies. The first two studies only reported accuracies without decimals. For this general performance, we used the same parts of the cases in the training data as Medvedeva et al. (2020). These are the procedure, the facts or both, depending on the article.

Discussion When we compare the accuracies of our models, trained and tested on the Medvedeva et al. (2020) dataset, we can see that our models perform similarly to those in the literature on the AC task. Note that the accuracies of SVM by Aletras et al. (2016) were obtained by training the SVM on parts of the case that were not available before the judgement was made, and this work should therefore be classified as outcome identification, rather than outcome prediction [1]. Performances on the BC task are lower than the ones achieved by the SVM, HIER-BERT and Legal-BERT, but higher than the one achieved by the regular BERT model. However, it is difficult to compare exact performances without using the exact same datasets. For example, BERT and HIER-BERT were trained on 7100 cases and their F1-score was calculated on predictions on unbalanced test sets (66% violation) [8], while we trained on 3133 cases and used balanced test sets. Across the four models that we use, there is no clear best model, and performance is dependent on the dataset, task and article.

Accuracies are generally slightly lower when training on the larger and more recent ECHR-OD dataset. We also see that the MCC ranks the performance of the models differently than accuracy does (except for Articles 3 and 5). The MCC more accurately depicts the actual performance of the models as it accounts for the rate of true positives, false positives, true negatives and false negatives, which leads to a more reasonable evaluation. The last two sets of rows in Table 2 therefore show that accuracies or F1-scores can show inflated results and can incorrectly suggest a high performance. Therefore, if we wish to take AI seriously in AI&Law, we should use more reliable metrics.

Investigating what Parts to use The cases from the ECHR consist of 6 different parts, including the judgement. Previous studies have used different parts of the cases to train their models, with mixed results. As discussed in the background section, only the introduction, procedure and facts are known before a case is tried. Some of the earlier literature has used parts that were made available after the case has been judged, however. If our goal is to take the law seriously, we should only include parts that are available before the judgement was made. We therefore do not use the law section, the judgement and the dissenting and concurring opinions. Additionally, the introduction, which contains only general information about a case will also not be used, as this should not have any predictive value. We therefore focus on the procedure, the facts, and a combination of both to look into which of these yields the best results.

Table 2
Comparison of the performance between various models and studies on the ECHR classification task. Results ($\times 100$) are accuracies, except those marked with $*$ (F1-scores), and the last four rows (Matthew's Correlation Coefficient).

Article	2	3	5	6	8	10	11	13	14	All
Results from literature										
SVM (Aletras et al., 2016)	-	75	-	**84**	**78**	-	-	-	-	-
SVM (Medvedeva et al., 2020)	73	**80**	71	80	72	61	**83**	**83**	75	-
SVM (Clavié and Alphonsus, 2021)	-	-	-	-	-	-	-	-	-	82.2*
BERT (Chalkidis et al., 2019)	-	-	-	-	-	-	-	-	-	17.0*
HIER-BERT (Chalkidis et al., 2019)	-	-	-	-	-	-	-	-	-	82.0*
Legal-BERT (Chalkidis et al., 2020)	-	-	-	-	-	-	-	-	-	**88.3***
Medvedeva et al. 2020 dataset										
SVM (replication)	72.6	**80.1**	**75.0**	80.7	70.6	63.7	82.5	**83.3**	75.8	67.6
NB	**77.0**	77.2	72.1	79.3	71.7	64.3	80.6	83.2	70.4	68.8
RF	71.8	70.9	77.3	68.2	73.6	**66.9**	61.8	80.0	**78.5**	69.6
BERT	70.8	74.5	70.0	70.5	63.54	59.9	75.0	79.2	76.4	69.9
ECHR-OD (Accuracy)										
SVM (replication)	68.0	64.5	65.6	76.8	68.1	61.6	65.4	80.4	76.6	69.3
NB	70.6	64.9	66.1	77.6	66.4	66.8	75.0	77.3	73.6	71.5
RF	66.2	60.8	62.8	74.6	61.8	66.1	75.6	77.1	74.4	72.6
BERT	67.1	62.0	64.9	72.0	61.4	61.1	65.2	72.0	74.0	71.2
ECHR-OD (MCC)										
SVM (replication)	38.8	36.8	29.0	31.1	53.8	36.1	23.1	31.0	61.0	53.2
NB	42.9	41.2	29.9	32.3	56.1	33.6	34.3	50.6	54.7	47.4
RF	45.3	32.5	21.6	25.6	49.2	24.1	32.2	52.4	54.8	48.8
BERT	42.4	34.9	24.1	30.0	44.3	22.8	23.1	34.0	45.1	48.7

We should note that all cases in the ECHR dataset were published after the cases were tried;their texts can therefore potentially contain implicit or explicit information that was not available before the case was tried, even in the introduction, procedure and facts sections [20].

To investigate which parts of the case are useful in court case prediction, we train each of the four model types (SVM, NB, RF, BERT) on the facts, procedure and both the facts and the procedure. We use the ECHR-OD dataset. We report the average MCC across a 10 fold cross validation for each classifier, trained on every individual article and all articles at once. This experiment expands upon the research done by Medvedeva et al. (2020) by exhaustively reporting the performances, in terms of MCC instead of accuracy, of each combination of our four models trained on all possible parts. This comparison can be seen in Table 3, where the best results for each classifier on a given article is shown in bold.

Discussion There is quite some variation between the MCC of models using different parts in Table 3. Determining which part to use is therefore important to obtain the highest possible performance. The facts and the combination of facts and procedure yield the best results across the combinations of part, model and article. The procedure alone ranks the worst. This means that the facts are an essential part when doing court case predictions. This is unsurprising, as this part contains all of the relevant information regarding the circumstances, background, applicant and relevant law from other documents. We see that adding the procedure can improve performance, but this is dependent on the combination of article and model. This supports the method used in [9], where different parts are used for each article. We also show that the performance is dependent on the combination of the parts used and the model used, and we base our conclusions on the MCC rather than the accuracy.

4.2. Experiment 2: Specialist vs. Generalist models

In previous research, models were either trained on each individual article (AC task) or on all articles at once (BC task); in the latter case, the model is tasked with predicting whether there has been any violation, regardless of what article was violated. The performance of our own models and models from the literature on this task can be seen in the rightmost column of Table 2. This approach can be compared to a human legal generalist, who has knowledge of all articles. Instead of just a single generalist, however, one could opt to use a team of legal specialists, where each person of the team is specialized in a different article. In this experiment, we examine these two different approaches to the BC task.

The first approach is to create a single Generalist model that is trained on all cases of the ECHR. In the second

Table 3
Comparison of MCC between various machine learning models across a 10 fold cross validation, when training on only the facts, only the procedure and both.

Article	**2**	**3**	**5**	**6**	**8**	**10**	**11**	**13**	**14**	**All**
Dataset size	462	1326	996	2066	1040	380	104	392	546	7376
SVM										
Procedure	31.64	25.3	24.52	43.77	19.82	22.25	30.98	51.53	38.1	38.75
Facts	33.07	**29.0**	31.14	**54.65**	**36.16**	**24.4**	48.09	54.18	48.0	**45.57**
Procedure + Facts	**36.8**	28.37	**35.35**	53.81	35.66	23.16	**55.78**	**61.0**	**53.15**	43.77
NB										
Procedure	34.25	22.34	31.35	41.75	27.92	32.11	50.6	51.15	39.61	42.91
Facts	30.48	**29.88**	32.25	50.96	33.55	33.95	56.28	54.09	**48.05**	48.2
Procedure + Facts	**41.15**	25.4	**34.97**	**56.08**	**37.82**	**34.31**	**67.2**	**54.68**	47.41	**49.72**
RF										
Procedure	20.09	27.31	22.71	41.02	21.87	31.34	52.43	47.15	36.83	45.27
Facts	**35.29**	21.63	25.63	48.79	24.08	29.19	**54.14**	**54.82**	43.12	**50.91**
Procedure + Facts	32.48	**28.3**	**32.74**	**49.21**	**26.92**	**32.18**	46.83	54.77	**48.81**	50.65
BERT										
Procedure	26.82	**27.92**	28.96	**48.73**	13.72	22.54	33.96	41.94	46.79	42.44
Facts	19.19	24.12	**30.0**	48.13	22.79	**29.46**	**39.92**	**57.99**	42.6	**45.53**
Procedure + Facts	**34.87**	24.16	28.14	44.26	**25.5**	23.09	32.55	45.09	**48.68**	43.92

approach, we train an Ensemble of models, wherein each model is specialized in a different article of the ECHR, akin to the team of legal specialists. Each model of this Ensemble is trained on identifying violations for just a single article, thus reducing the problem space and potentially increasing performance. Additionally, such an Ensemble would be able to tell what article was violated, thus providing explanations for its decisions. Each model of the Ensemble would, however, have less data than the Generalist model, which might decrease performance. We perform an experiment to determine which approach yields the best performing model. We create three types of datasets:

- nine **Ensemble Training Sets**, one for each article, containing 90% of all of the cases that consider that specific article; the features are the facts of the case, and the label is whether or not there is a violation of the respective article in the case.
- the **Generalist Training Set** contains all of the cases from all nine Ensemble Training Sets; the features are again the facts, and the label is whether or not *any* article was violated in a case.
- the **Testing Set** contains the 10% of cases not used in the Ensemble Training Set and Generalist Training Set; features and label are the same as in the latter.

We generate the 90% - 10% split in training and test data randomly, preserving the 50% balance in classes. We use these three datasets to train and test a Generalist model and an Ensemble. The **Generalist model** is trained on the Generalist Training Set and evaluated using the Testing Set. The **Ensemble** consists of nine specialist models, each trained on a different Ensemble Training Set. Each of these specialist models will be tasked with predicting the labels of the cases from the Test Set. The predictions of each specialist model will be combined in a disjunctive manner to form the final prediction of the Ensemble. In other words, the output will be violation if *any* specialist model predicts violation, and non-violation otherwise.

Table 4
MCC for Generalist models and (improved) Ensembles.

	General	Ensemble	Improved Ensemble
SVM	**37.04**	27.55	31.99
NB	30.33	**30.54**	5.42
RF	**35.95**	14.58	17.43
BERT	**27.66**	13.51	24.95

We compare the performance of the Generalist model to that of the Ensemble. The experiment is performed for every one of our four model types: the SVM, NB, RF and BERT. We also repeat every experiment 10 times for each type of model, using different cases for the training and testing sets in each iteration. We report the average MCC in Table 4. The best results are shown in bold.

Discussion From Table 4 we see that for most Generalist models the MCC is much higher than that of the Ensemble. The Generalist models therefore outperform the Ensembles for most types of classifiers. The exception is the NB classifier, where there is little difference in MCC. This suggests that, in predicting ECHR court cases, a larger problem space combined with more training data results in better performance than a reduced

Table 5
Confusion matrix of all predictions by the SVM Ensemble.

	Predicted:		
True:	**N-violation**	**Violation**	**Total**
N-violation	623	1,138	1,761
Violation	1,138	17,601	18,739
Total	1,761	18,739	

Table 6
Confusion matrix of all predictions by the Improved SVM Ensemble.

	Predicted:		
True:	**N-violation**	**Violation**	**Total**
N-violation	622	1,139	1,761
Violation	934	17,805	18,739
Total	1,556	18,944	

problem space with less data.

Each of the specialist models in an Ensemble is trained on cases that pertain to a single article. The Test Set used for these specialist models, however, considers all articles, most of which the individual specialist models will not have seen during training. Ideally, if a specialist models is presented with a case that considers an article that it is not trained for, it should predict 'non-violation'. However, it is not explicitly trained to give that prediction and, as a result, provides a random prediction. Given that the final prediction of the Ensembles is an OR-function, this leads to many incorrect 'violation' predictions. This can be seen in Table 5, which displays the confusion matrix of all of the predictions done by the SVM Ensemble. The Ensemble predicted 'violation' in a total of 18,739 cases, out of which 17,601 were correct and it predicted 'non-violation' in a total of 1,761 out of which only 623 were correct. This initially might seem like a great performance, and would lead to an accuracy of 89.9%. However, the Test Set is heavily skewed towards violation (91.14% of all cases). The performance on violation cases is therefore relatively good, accurately predicting the outcome of 93.9% of all violation cases. However, the model also incorrectly assigns violation to 64.6% of all non-violation cases.

Improved Ensemble A potential solution to this issue is to present the specialist models with cases that do not pertain to the article that they are focused on during training. For example, a specialist model trained on article 6 cases could also explicitly be trained to predict 'non-violation' for all cases that do not pertain to article 6. To create this new Improved Ensemble, we alter the training datasets as follows. We create nine **Improved Ensemble Training Sets**, one for each article. Just as the earlier Ensemble Training Set, each Improved Ensemble Training Set contains 90% of all of the cases that consider that specific article. The features in these datasets are the facts of the case, and the labels are whether or not there is a violation of the respective article in the case. Additionally, we add cases from other articles to this dataset, where each additional case has the 'non-violation' label. Since almost all articles contain more violation than non-violation cases (see Table 1), we add these additional cases to each Improved Ensemble Training Set until their number of violation and non-violation cases is equal. The **Improved Ensemble** is set up in the same way as the earlier Ensemble, but each specialist model of this Improved Ensemble is trained on the Improved Ensemble Training Sets. The results of the Improved Ensemble are shown in the rightmost column of Table 4.

Discussion The Improved Ensemble performs better than the initial Ensemble when using the SVM, RF and BERT models. This supports our idea that the specialist models generally perform better when including additional cases from other articles with a 'non-violation' label. This informs the model to predict 'non-violation' for cases pertaining to other articles.

The NB models, however, seem to perform worse in the 'Improved Ensemble' scenario. Analysing the data shows that the specialist models in the 'Improved' NB Ensemble now predict 'violation' in 99.5% of all test cases, thus yielding a low MCC in Table 4. Our hypothesis is that the problem space might have become too large for the NB specialist models by providing them with a relatively small number of additional cases pertaining to other articles. Investigating this idea is left for future research. Note that the accuracy of this Ensemble is 91.2%, incorrectly suggesting a high performance of this irresponsible system, which further advocates for the use of the MCC as a performance metric.

Table 4 shows that the Improved SVM Ensemble performs better than the initial SVM Ensemble. The confusion matrix of this Improved SVM Ensemble is shown in Table 6. Here, we see that the performance on the non-violation cases is almost identical to that of the initial Ensemble on the same cases, as shown in Table 5. By including the additional cases, our aim was to instruct the specialist models to predict non-violation for cases that did not pertain to its specific article. However, while the Improved Ensemble does perform better, Table 6 shows us that the Improved Ensemble did not improve its performance on the non-violation cases. The difference in performance is therefore due to the Improved Ensemble's predictions on the violation cases. We see that the Improved Ensemble correctly predicts the outcome of 95% of all violation cases. This is 1.1 percent point higher than the performance of the initial Ensemble.

By including additional cases pertaining to other articles in each of the specialist models, we are able to somewhat improve the Ensemble's performance in most

cases. It should be noted here that we only include a small subset of additional cases pertaining to other articles. The number of additional cases equals the number of violation cases of a given article, minus the number of the non-violation cases of that article. This is to ensure a 50% violation rate in the training dataset of each model. The Ensemble could potentially be improved further by oversampling violation cases for each article and including more of these additional cases. This could potentially change the results of the 'Improved' NB Ensemble as well, as it would then have more cases per article. However, oversampling has downsides as well, such as overfitting. Future research could investigate this idea.

4.3. Experiment 3: Temporal effects

In the previous experiments, we have split our training and testing sets randomly, or used 10-fold cross validation to train and test our models, just as much previous research has done. By splitting the data randomly, we might select training cases that occurred more recently than our test cases. In other words, we might use future cases to predict cases from the past. When it comes to court case predictions, an argument can be made for selecting the most recent cases as the test set and to use the older cases as a training set. This way, we use past cases to forecast the future. Some models in previous research were trained in this way [8], but effects of this design choice have not yet been studied. To evaluate how models perform under these different temporal circumstances, we narrow our scope and focus on article 6 of the ECHR. This article contains the most cases and has therefore been investigated in other work as well [**?** 11]. In this experiment, we use the facts of the cases from article 6 as the features, and whether or not article 6 was violated as the label. We train models on cases from the past and evaluate how their performance compares to models trained on randomly selected data. We use three types of datasets in this experiment, each containing only cases from article 6:

- test sets consist of cases from a single **Test Year**.
- the models trained on **Past Cases** will be trained on all cases that occurred before the Test Year.
- the models trained on **Random Cases** are trained on randomly selected cases that occurred either before or after the Test Year.

For each year, we therefore generate 2 types of models, one trained on Past Cases (cases from years before the year that we use to test the model) and one trained on Random Cases. Note that we ensure that the size of the Random Cases training set is the same as the size of the Past Cases training set for each respective Test Year. This way, we can disregard discrepancies in the size of the training dataset as a potential cause of differences in performance. All training sets are balanced, such that half of the cases are violation cases and the other half are non-violation cases. We use all four model types (SVM, NB, RF, BERT) and train on the facts of the ECHR-OD dataset. The results can be seen in Table 7.

Table 7
Mean MCC of models predicting all cases from a given year, after training on cases from the past or training on a random sample of a similar size. We display mean results across all Test Years (1979-2022) and results across more recent Test Years (2000-2022).

	1979 - 2022		**2000-2022**	
	Random	Past	Random	Past
SVM	**19.60**	18.62	**35.32**	34.98
NB	16.40	**19.68**	29.38	**32.07**
RF	**21.11**	13.39	**33.38**	30.16
BERT	**21.78**	8.16	**26.68**	26.30

Discussion We performed the experiment using all possible Test Years from 1979 until 2022. Across all these Test Years, we see in Table 7 that the MCC of models trained on random cases is generally higher than the MCC of models trained on past cases. This would imply that learning from past cases is more difficult than learning from random cases from both past and future. The exception here is the NB classifier, which performs better when trained on past cases. In Figure 2, we plot the MCC per Test Year for each of the four model types trained on either Past Cases (blue) or Random Cases (orange). Note that the y-axis of each subplot is scaled differently. Here we see that the MCC of models trained on both random cases and past cases fluctuates a lot for earlier Test Years. This could be due to the limited number of available cases in those years that the models are used to train on (see the case distribution per year in Figure 1). The difference between the two (the shaded area in Figure 2) also decreases with more recent Test Years.

If we look only at the recent years (2000-2022) in Table 7, we get a more nuanced comparison. Table 7 still shows higher MCCs for models trained on random cases over models trained on past cases, with the exception of the NB model, but the difference between the two is much lower. The absolute mean MCC is also higher across the years. While the differences between the two approaches may be smaller, they still exist, as seen in Figure 2. Not only is it legally more reasonable to train on past cases to predict future cases, a random split of the data into a train and test set can also have an impact on the performance. If we wish to take the law more seriously in this type of research, to ensure realistic results we should train on past cases and test on future cases.

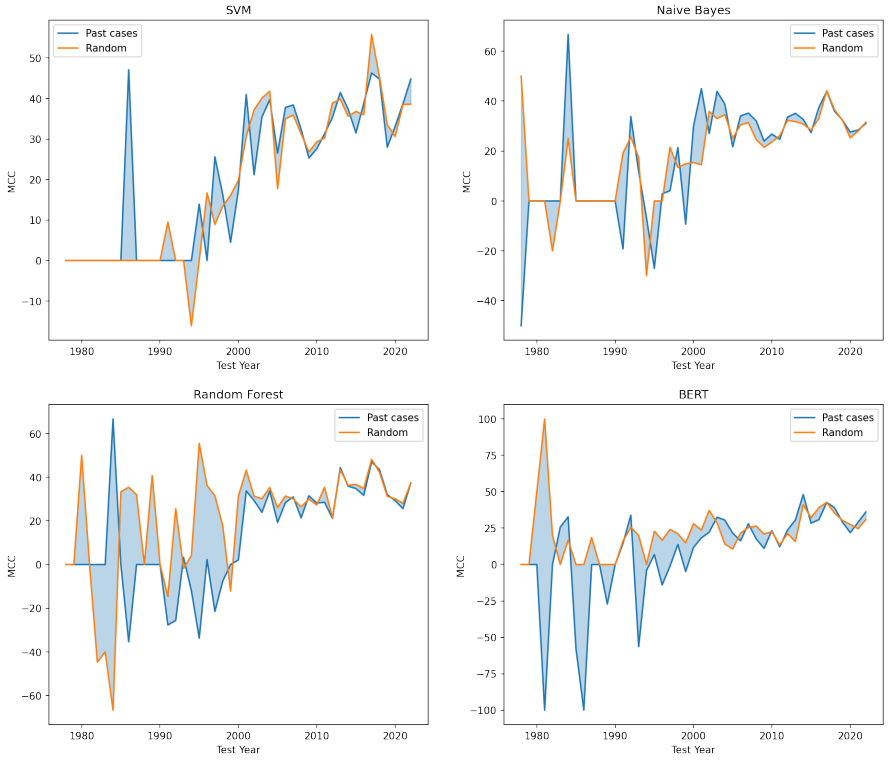

Figure 2: MCC of models trained on either cases from the past (blue) or randomly selected cases (orange) and tested on cases of a given Test Year, ranging from 1978 to 2022.

Time Window We know that machine learning systems tend to perform better with more data, granted that the data is a proper reflection of the problem space. We also know, however, that the interpretation of the law is subject to change over time and precedent may be overturned. When this happens, the older cases can be overruled by newer cases. Looking too far into the past may therefore not be optimal when trying to predict new court cases. We therefore also investigate the difference between using all cases or only a subset of the more recent cases.

We train our models on cases from the past but only on a limited number of recent years. This window of years represents how far we look back into the past, and will be varied between 5 and 35 years in steps of 5. We test on only a single Test Year of cases. For example, training the model to predict cases from 2022 using a window of 5 years means that the model is trained on the cases from 2016 up to and including 2021. Because we only have cases from 1978 until 2022, we will test our models on cases from 2013 to 2022. This way, we can train our model using windows of up to 35 years for each given Test Year.

We create a model for each combination of type (SVM, NB, RF and BERT), Test Year (from 2013 to 2022), and window of training data (5 to 35 in steps of 5). There are therefore $4 * 10 * 7 = 280$ different setups in this experiment. To represent the results, we average the MCC of each of these models across the years, as to show the effects of the window size on the performance of each of the model types. We train on the facts of cases from Article 6 and use the ECHR-OD dataset. The results of this experiment are shown in Table 8, where the best window for each model is shown in bold. The last column in Table 8 reports the standard deviation across the different windows.

Table 8
Mean MCC of the window experiment.

Window	**5**	**10**	**15**	**20**	**25**	**30**	**35**	**SD**
SVM	37.5	37.5	37.4	38.4	**39.9**	39.4	39.6	1.1
NB	32.0	30.8	31.4	31.7	32.2	31.9	**32.4**	0.5
RF	36.6	**36.8**	35.4	34.8	35.0	35.0	35.4	0.8
BERT	33.3	29.6	31.0	27.5	30.7	31.8	**34.3**	2.3

Discussion Based on Table 8 we cannot extrapolate a clear relationship between the number of past years of cases in training and the performance of the models. The NB and BERT models perform best with 35 years worth of cases. However, there does not seem to be a clear positive relationship between the window size and MCC, as the MCC increases and decreases slightly across the window sizes. A similar observation holds for the SVM and RF models, which do not perform much better or worse with more years worth of cases from the past. The overall impact of the window is therefore small, as also indicated by the low standard deviations. The BERT model is impacted most by the window size, but the standard deviation in MCC is still only 2.3. In this scenario, using cases from further in the past does therefore not seem to have a significant impact on the performance of the models. This can, of course, be different if the interpretation of the law has changed significantly. If we want to take the law more seriously, we should investigate the temporal effects of the legislation and, if precedent is overturned, adjust and evaluate our training data accordingly.

5. Conclusion

The approaches to court case predictions are diverse and difficult to compare [1]. While some methods yield better results, they may also raise concerns about how reasonably they align with the characteristics of legal decision-making. For a proper analysis of court case prediction research, we should consider the unique characteristics of the law and the effects that it can have on the models. While justification, explainability and responsibility are major issues in machine learning and law, our scope did not include these aspects and focused instead on design choices.

If we want to take the law more seriously in machine learning research, we should measure the effect of relevant design choices and effects. We therefore propose to use the Matthew's Correlation Coefficient rather than the accuracy or F1-score, as the latter two metrics tend to yield inflated results and can incorrectly attribute a much higher performance to a model.

Based on our results of Experiment 1, the facts are the most important of a case when it comes to court case predictions (see Table 3). Including the procedure of the case can increase performance, but this is dependent on the article and on the model used. For the best results, the parts used should therefore be included in the parameter optimization pipeline.

In Experiment 2, our Generalist model, trained on all articles at once, outperforms our Ensemble of specialist models each trained on a specific article. While the specialist models had a reduced problem space, more data appeared to still be more important. By including some additional cases of other articles in the training phase of each specialist model of the Ensemble, we are able to increase performance for most model types. While we only included a small sample of these additional cases, future research could investigate whether including more additional cases in the training data of specialist models of the Ensemble could increase performance further.

Experiment 3 shows that training on past cases to predict the future is more difficult than training on randomly selected instances from both the past and the future. Taking into account the effects of time may therefore have an effect on performance, especially when you consider much older cases. For that reason, randomly splitting data into a training and test set, or running a k-fold cross validation, might show unrealistic results. In these scenario's, it is more than likely that the model is predicting past cases using future cases, which is impossible in reality and does not account for the temporal aspects of the law.

We show that using only a limited number of years worth of cases, rather than all cases, does not seem to have an impact on the performance of our models, as shown in Table 8. This suggests that the interpretation of the law, in particular regarding article 6 of the European Court of Human Rights, remained stable enough for machine learning predictions. There are, however, legal considerations that might suggest the removal of certain older cases, especially after certain landmark cases or changes in society. In those cases, we should investigate the effects that this has on the legislation and adjust our training data accordingly.

We have explored legally reasonable design choices and effects in court case predictions, and have shown their impact on performance. We conclude that, taking the law more seriously in machine learning research requires that the relevant, unique characteristics of the law are taken into account. Our findings are by no means enough to address inherent limitations (in particular with respect to justification), and future research has to remain critical of the choices that are being made in order to remain legally reasonable.

Acknowledgments

This research was funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl.

References

[1] M. Medvedeva, M. Wieling, M. Vols, Rethinking the field of automatic prediction of court decisions, Artificial Intelligence and Law (2022) 1–18.

[2] T. Bench-Capon, The need for good old fashioned AI and Law, International Trends in Legal Informatics: A Festschrift for Erich Schweighofer (2020) 22–36.

[3] C. Steging, S. Renooij, B. Verheij, Discovering the rationale of decisions: towards a method for aligning learning and reasoning, in: J. Maranhão, A. Z. Wyner (Eds.), ICAIL '21: Eighteenth International Conference for Artificial Intelligence and Law, São Paulo Brazil, June 21 - 25, 2021, ACM, 2021, pp. 235–239.

[4] T. Santosh, S. Xu, O. Ichim, M. Grabmair, Deconfounding legal judgment prediction for European Court of Human Rights cases: Towards better alignment with experts, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 1120–1138. URL: https://aclanthology.org/2022.emnlp-main.74.

[5] I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. Katz, N. Aletras, LexGLUE: A benchmark dataset for legal language understanding in English, 2021. arXiv:2110.00976.

[6] A. Quemy, R. Wrembel, ECHR-OD: On building an integrated open repository of legal documents for machine learning applications, Information Systems 106 (2022) 101822.

[7] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, V. Lampos, Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective, PeerJ Computer Science 2 (2016) e93.

[8] I. Chalkidis, I. Androutsopoulos, N. Aletras, Neural legal judgment prediction in English, 2019. arXiv:1906.02059.

[9] M. Medvedeva, M. Vols, M. Wieling, Using machine learning to predict decisions of the European Court of Human Rights, Artificial Intelligence and Law 28 (2020) 237–266.

[10] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: The muppets straight out of law school, ArXiv preprint arXiv:2010.02559 (2020).

[11] J. Mumford, K. Atkinson, T. Bench-Capon, Reasoning with legal cases: A hybrid ADF-ML approach, in: Legal Knowledge and Information Systems, volume 362 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2022, pp. 93–102.

[12] L. Zheng, N. Guha, B. Anderson, P. Henderson, D. Ho, When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset of 53,000+ Legal Holdings, Association for Computing Machinery, New York, NY, USA, 2021, p. 159–168.

[13] B. Clavié, M. Alphonsus, The unreasonable effectiveness of the baseline: Discussing SVMs in legal text classification, in: S. Erich (Ed.), Legal Knowledge and Information Systems - JURIX 2021: The Thirty-fourth Annual Conference, Vilnius, Lithuania, 8-10 December 2021, volume 346 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2021, pp. 58–61.

[14] T. Pranckevičius, V. Marcinkevičius, Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification, Baltic Journal of Modern Computing 5 (2017) 221.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Brew, Huggingface's transformers: State-of-the-art natural language processing, CoRR abs/1910.03771 (2019).

[17] S. Bird, E. Klein, E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit, " O'Reilly Media, Inc.", 2009.

[18] B. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochimica et Biophysica Acta (BBA)-Protein Structure 405 (1975) 442–451.

[19] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genomics 21 (2020) 1–13.

[20] M. Medvedeva, A. Üstün, X. Xu, M. Vols, M. Wieling, Automatic judgement forecasting for pending applications of the European Court of Human Rights., in: ASAIL/LegalAIIA@ ICAIL, 2021, pp. 12–23.