Predicting Supreme Court Decisions

Anonymous

1 Introduction

In the United States Supreme Court, appealed cases are heard and discussed by six to nine judges, who can either uphold the previous decision (unsuccessful appeal) or reverse it (successful appeal). These decisions provide valuable insights into legal system in the US, shaped by a range of factors. Nonetheless, much of the existing research focuses primarily on factors such as the personal attributes of justices, rather than examining the procedural elements of the hearings themselves. To address the gap, we employ the Super-SCOTUS dataset (Fang et al., 2023b) to answer the following questions:

RQ1. Does the information available prior to or during hearing help predict Supreme Court decisions?

RQ2. Do the features reflecting the difficulty of cases contribute to prediction of Supreme Court decisions?

RQ3. How does gender of justices affect prediction of accuracy, potentially indicating bias?

In response to these questions, we propose the following hypothesis:

H1. The information available prior to or during hearing does help predict Supreme Court decisions

H2. Incorporating features reflecting the difficulty of cases improves model performance in predicting Supreme Court decisions.

H3. The gender of justices influences Supreme Court decisions.

By testing these hypotheses, we aim to better understand how various factors influence Supreme Court outcomes.

2 Related Work

Lim (2000) analysed 600 relations of Supreme Court cases between precedent and later decisions, concluding that it is statistically significant that justices' past decisions influence their present decisions in similar cases. Gleason (2024) found that female attorneys may be less successful at the U.S. Supreme Court under certain conditions, particularly due to gender norms and implicit biases about expected behaviours for men and women. As discussed, there is limited research investigating how procedural elements contribute to the prediction of Supreme Court decisions. To address the gap, this study utilizes a comprehensive dataset (Fang et al., 2023b) that connects different sources at various stages of the court hearing to investigate how various factors help predict Supreme Court decisions.

3 Method

3.1 Feature Selection

For RQ1, a Chi-squared test filters categorical features, avoiding unnecessary one-hot encoding that causes high dimensionality and inefficient training. Referring to Table 1, only features with a p-value below 0.01 are included. The low p-values may arise from high degree of freedom or presence of many unknown values, reducing their predictive power. After one-hot encoding and embedding, an ANOVA F-test selects the top 200 features, improving generalization and avoiding the curse of dimensionality (Chen et al., 2020).
Additionally, a new feature, *argument month*, is engineered from the *argument date*. Argument year is not used to avoid redundant correlation with *year*, and argument day is excluded due to high variability.
For RQ2, a new feature, *length of hearing* (calculated as the difference between argument and decision dates) is introduced alongside the top 200 features.
For RQ3, the percentage of female justices, is added to the 200 selected features.

Feature	Chi-Square	P-value	Degrees of Freedom
Title	4541.8103	4.5622e -01	4532
Petitioner	3573.0834	3.1720e -02	3418
Respondent	3371.8951	4.6330e-01	3365
Petitioner State	50.2199	6.9239e-01	56
Respondent State	66.2775	1.4184e-01	55
Petitioner Category	409.8353	2.5659e-10	246
Respondent Category	396.6929	2.2668e-11	227
Issue area	62.2141	4.7807e-08	14

Table 1- The result of Chi-squared tests on categorical data

3.2 Data Preprocessing

For categorical features, one-hot encoding is employed, ignoring unknown categories during transform. Normalization, specifically MinMax, minimizes bias from features with higher numerical contributions, enhancing model performance (Singh & Singh, 2020).

After feature selection, missing values in *argument month* are filled with the most frequent month, while missing values in *length of hearing* are filled with the mean. This approach preserves feature integrity without introducing significant bias due to low percentage of missing data.

Figure 2 shows a slight class imbalance in the dataset, hence no adjustment is performed. Mild imbalances typically do not adversely affect model performance, instead adjusting them through oversampling or undersampling may lead to overfitting or loss of information (Thölke et al., 2023).

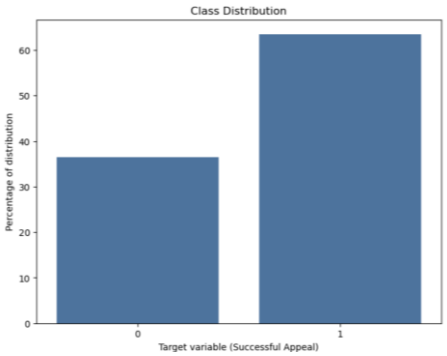

Figure 2- Class distribution of target variable (successful appeal)

3.3 Machine Learning Model

The Zero-R model, which uses the most frequent label serves as a baseline for evaluating the performance of more complex models in this study: K-Nearest Neighbours (KNN) and Logistic regression.

KNN computes the distance between a test instance and all training data points to identify the K closest neighbours, determining the labels based on these neighbours (Uddin, 2022). While, Logistic Regression predicts a binomial outcome, providing probabilities based on input variable (Ray, 2019). Table 4 summarises the strengths, weaknesses, advantages and disadvantages of the models. KNN and Logistic Regression are chosen for their strength to handle the characteristic of data well. Besides, they provide a comparison between the performance of non-linear and linear models.

	KNN	Logistic Regression
Strength	• Effectively handle unseen categories • Effectively handle mixed feature types • Capture non-linear relationships	• Can handle mixed feature types, when encoded appropriately • Capture linear relationships • Resilient to small noise and multicollinearity
Weakness	• Struggle with high-dimensional data • Irrelevant features can degrade accuracy	• Ineffectively handle unseen categories • Require all independent variables to be identified to perform well
Advantage	• Simple to implement • Cost-effective to build	• Simple to implement • Computationally efficient
Disadvantage	• Require scaling • Computationally intensive for large dataset	• Prone to overfitting

Table 3- Strengths and weaknesses, advantages and disadvantages of KNN and Logistic Regression

3.4 Hyperparameters Search

Hyperparameter search involves identifying, optimizing, and comparing hyperparameters. Table 4 summarizes the identified hyperparameters for optimization.

Using grid search, all combinations of the specified hyperparameters are evaluated to identify the best combination for predicting decision based on accuracy and f1 score. The search is conducted on validation set (development set), allowing us to identify the optimal hyperparameters for each metric. The tuned models are then trained on the combined training and validation sets. The refined models are evaluated using the Kaggle competition results, with the model achieving highest accuracy selected as the final model. Table 5 summarises the best hyperparameters which maximises the model performance.

The best hyperparameters for KNN and Logistic Regression will be used to investigate RQ2 and RQ3.

	Hyperparameters	Options
KNN	Distance Metric	Euclidean, Manhattan, Hamming
	Number of neighbours, k	20 ~ 50
	Weighting method	Uniform, Distance
Logistic Regression	Maximum iterations	1 ~ 100
	Solver type	liblinear, lbfgq or newton-cholesky
	Regularization parameter, c	0.1, 1, 10

Table 4- Hyperparameters for optimization in KNN and Logistic Regression

		Best Hyperparameters using …	
		Accuracy	F1 score
KNN	Distance metric	Euclidean	Hamming
	Number of neighbours, k	20	57
	Weighting method	Distance	Uniform
	Accuracy (validation set)	0.6638	0.6464
	F1 (validation set)	0.7734	0.7811
	Accuracy (test set)	0.7011	0.6609
Logistic Regression	Maximum iterations	39	6
	Solver type	lbfgs	lbfgs
	Regularization parameter, c	1	0.01
	Accuracy (validation set)	0.6534	0.6412
	F1 score (validation set)	0.7590	0.7800
	Accuracy test set)	0.6724	0.6322

Table 5- Comparison of results for KNN and Logistic Regression with best hyperparameters for maximizing accuracy and f1 score in RQ1

3.5 Evaluation metrics

This study primarily evaluates the performance using accuracy since the primary goal is to predict Supreme Court decisions more accurately and accuracy is the only metric available for test set. The F1 score is considered due to the mild class imbalance in the dataset, providing more informative insights (Thölke et al., 2023). Additionally, bias and variance are assessed through learning curves, which plot accuracy against the number of neighbours (for KNN) or iterations (for Logistic Regression). Bias help detect potential underfitting and ensure that the model does not favour the majority class. While, variance can identify possible overfitting, reducing generalization.

4 Results

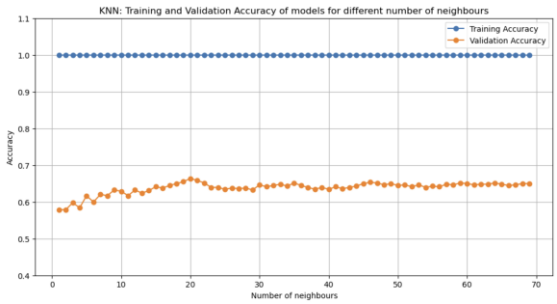

Figure 6- Learning curve for KNN, where weighting method is distance and distance metric is Euclidean

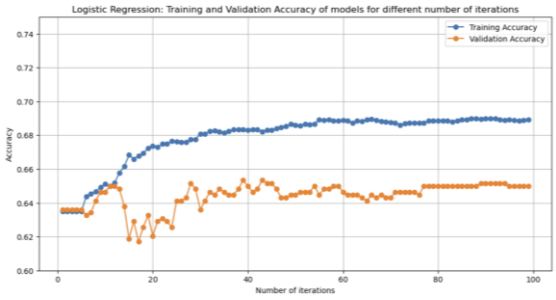

Figure 7- Learning curve for Logistic Regression, where solver typer is lbfgq and regularization parameter is 1

		Training set	Validation set	Test set
Baseline model	*Accuracy*	*0.6351*	*0.6360*	*0.6322*
	F1 score	*0.7768*	*0.7775*	-
KNN	*Accuracy*	*1*	*0.6638*	*0.7011*
	F1 score	*1*	*0.7734*	-
Logistic Regression	*Accuracy*	*0.6464*	*0.6534*	*0.6724*
	F1 score	*0.7811*	*0.7590*	-

Figure 8- Comparison of accuracy and f1 score for baseline model, KNN and Logistic Regression with best hyperparameters across different training, validation and test sets for RQ1

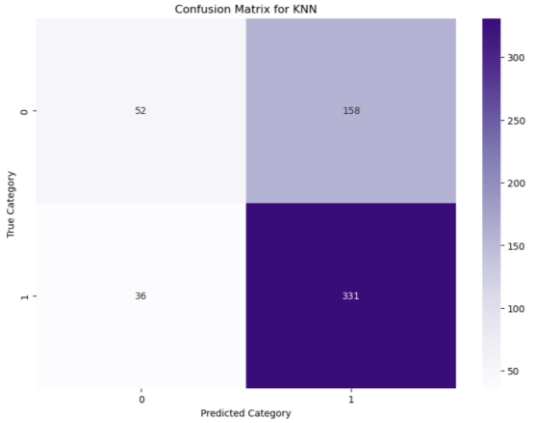

Figure 9- Confusion matrix for KNN with best hyperparameter on validation set for RQ1

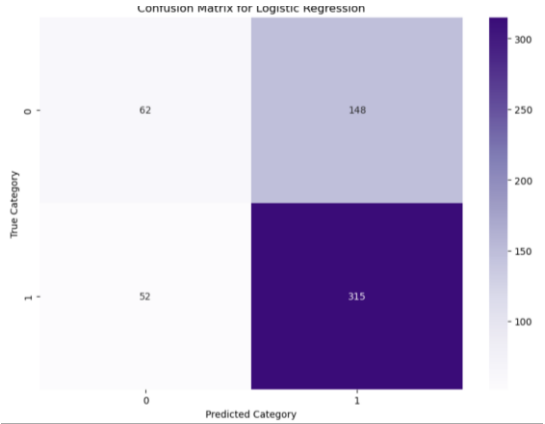

Figure 10- Confusion matrix for Logistic Regression with best hyperparameter on validation set for RQ1

	KNN		Logistic Regression	
	Validation	Test set	Validation	Test set

	set		set	
RQ1	0.6638	0.7011	0.6534	0.6724
RQ2	0.6551	0.7126	0.6412	0.6782
RQ3	0.6326	0.7011	0.6482	0.6609

Table 11- Accuracy of KNN and Logistic Regression for validation and test set across different RQs

	Training set	Validation set
Correlation	-0.0046	0.0306
p-value	0.7523	0.4633

Table 12- Correlation between hearing length and target variable on training and validation set

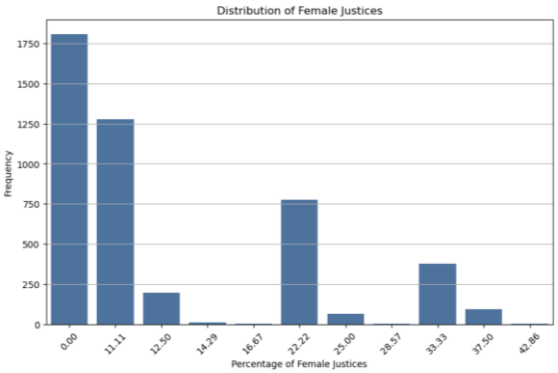

Table 13- Distribution of female justices in the dataset

5 Discussion

H1. The information available prior to or during hearing does help predict Supreme Court decisions

Based on Table 9, both KNN and Logistic regression have higher accuracy than the baseline model across training, validation and test sets. This indicates the information available prior or during court hearing are useful in predicting the decisions. KNN's higher accuracy suggests there are non-linear and complex relationship among some variables. This is consistent with the study by Martin et al. (2004), where non-linear model is more accurate in predicting court decision due to their ability to capture more complex relationship. In the validation set, KNN achieves a similar F1 score to the baseline model, while Logistic Regression is lower, suggesting that KNN balances precision and recall better. As shown in Figures 10 and 11, KNN's higher F1 score is due to higher recall (more predictions of positive classes), but it comes at the expenses of precision (more false positives).

Figure 7 illustrates that validation accuracies converge for KNN, suggesting no bias. Moreover, despite the large gap between testing and validation accuracies, KNN performs well on the test set, suggesting that it generalizes effectively and does not overfit. The large gap is caused by the testing accuracies which remain at 100% as the number of neighbours increase. This 100% test accuracies are due to the distance weighting function in KNN. According to the scikit-learn documentation (Garreta, 2013), the weight is calculated as the inverse of the distance between data points. When the training set is used for testing, the closest neighbour to each point is the point itself, resulting in a distance of 0 and creates an infinite weight for the closest point, leading to correct predictions in all cases.

For Logistic Regression, Figure 8 illustrates that both testing and validation accuracies converge, indicating no underfitting. The insignificant gap between testing and validation accuracies also suggests no overfitting.

In short, both models which incorporate information prior to or during the hearing outperform the baseline model, demonstrating good generalisation to unseen data. H1 is supported, answering RQ1 that the information prior to or during hearing are useful in predicting Supreme Court rulings.

However, their accuracies (65%-70%) reflects the complexity of predicting Supreme Court decision. Further research suggests more advanced models, such as Hassanaat KNN (Uddin et al., 2023) and SVM (Rosili et al., 2021) could improve accuracy beyond classic KNN.

H2. Incorporating features reflecting the difficulty of cases improves model performance in predicting Supreme Court decisions.

Table 11 shows that adding the feature, *length of hearing*, which reflects the complexity of cases, slightly reduce KNN and Logistic Regression performance on the validation set. This suggests the feature maybe irrelevant and introduces noise. This may be because the hearing length affected by factors beyond case difficulty such as the requirement for decisions to be released by the summer recess (Supreme Court of the United States, n.d.). Besides, Table 12 shows that hearing length shows no strong correlation with *successful appeal*, with inconsistent association between the training and validation sets and insignificant p-values.

However, both models show a slight improvement in performance on the test set (Table 11). This may be a coincidental, but we do not entirely rule out the possibility that *length of hearing* may still offer some predictive value, despite its weak correlation with the outcome (Kumar & Chong, 2018). More unseen data is required to determine whether the

improvement is coincidental or meaningful.
The inconsistent performance across datasets suggest H2 is not supported, answering RQ2: Case difficulty is not helpful in predicting Supreme Court decisions.

H3 The gender of justices influences Supreme Court decisions.

Table 11 shows the feature introduced, *percentage of female justices* degrades or maintains model performance on validation and test sets. This contradicts existing research which show that gender does affect court decision, indicating potential bias. For instance, Jeniffer (2005) proves that female judges increase the likelihood of supporting plaintiff, and Songer & Crews-Meyer (2000) found that female judges are more likely to support liberal position.
This contradiction may arise from the dataset distribution, as illustrated in Figure 13. Around 65% of cases contains no or only one female justice, raising concerns about the models' ability to detect any influences of gender. The skewed representation of female justices may prevent the models from capturing the relationship between gender and successful appeals due to insufficient data points for analysis.
Nonetheless, most research on gender effect gender in court decisions is dated. The trend may have changed over time, gender may no longer have a significant effect on court decisions. More recent data with higher representation of female justices is necessary to further investigate the effect of gender on court decisions.
In brief, since the introduced feature, *percentage of female justices* does not improve the performance of models, H3 is not supported. This answers RQ3: the gender of justices does not significantly affect the Supreme Court decisions.

6 Conclusions

This study explores how information available prior to or during hearings can predict Supreme Court decisions. The analysis demonstrates that these factors are useful for predicting rulings, but more advanced machine learning algorithms are needed to improve accuracy.
The investigation of *length of hearing* reveals that case difficulty does not enhance model performance, primarily due to conflicting results between validation and test sets. More unseen datasets are necessary for conclusive findings. Moreover, the *percentage of female justices* does not improve prediction accuracy, contradicting previous research that suggests gender influences judicial decisions. The limited representation of female judges (65% of cases involve no female judges or only one) likely affects the models' ability to assess gender impact. More recent cases with balanced gender representation are required to further explore this.
Overall, this study provides valuable insights into judicial decision-making and highlight the need for additional data and research to enhance our understanding of the U.S. legal system.

References

Fang, B., Cohn, T., Baldwin, T., and Frermann, L. (2023b). Super-SCOTUS: A multi-sourced dataset for the Supreme Court of the US. In Preot,iuc-Pietro, D., Goanta, C., Chalkidis, I., Barrett, L., Spanakis, G., and Aletras, N., editors, *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 202 – 214, Singapore. Association for Computational Linguistics.

Garreta, R. (2013). *Learning scikit-learn : Machine Learning in Python*. Packt Publishing.

Gleason, S. A. (2024). I Can't See You; Can You Hear Me? Gender Norms and Context During In-Person and Teleconference US Supreme Court Oral Arguments. *Politics & Gender*, *20*(2), 318-345.

Jennifer L. P. (2005). Female Judges Matter: Gender and Collegial Decision making in the Federal Appellate Courts. *The Yale Law Journal*, *114*(7), 1759–1790.

Kumar, S., & Chong, I. (2018). Correlation Analysis to Identify the Effective Data in Machine Learning: Prediction of Depressive Disorder and Emotion States. *International Journal of Environmental Research and Public Health*, *15*(12). https://doi.org/10.3390/ijerph15122907

Lim, Y. (2000). An Empirical Analysis of Supreme Court Justices' Decision Making. *The Journal of Legal Studies*, *29*(2), 721–752. https://doi.org/10.1086/468091

Martin, A. D., Quinn, K. M., Ruger, T. W., & Kim, P. T. (2004). Competing Approaches to Predicting Supreme Court Decision Making.

Perspectives on Politics, *2*(4), 761–767.

Ray, S. (2019). A Quick Review of Machine Learning Algorithms. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019 International Conference On*, 35–39. https://doi.org/10.1109/COMITCon.2019.8862451

Rosili, N. A. K., Zakaria, N. H., Hassan, R., Kasim, S., Rose, F. Z. C., & Sutikno, T. (2021). A systematic literature review of machine learning methods in predicting court decisions. *IAES International Journal of Artificial Intelligence*, *10*(4), 1091–1102. https://doi.org/10.11591/IJAI.V10.I4.PP1091-1102

Rung-Ching Chen *et al.* Selecting critical features for data classification based on machine learning methods. Journal of Big Data, *[s. l.]*, v. 7, n. 1, p. 1–26, 2020.

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing Journal*, *97*. https://doi.org/10.1016/j.asoc.2019.105524

Songer, D. R., & Crews-Meyer, K. A. (2000). Does Judge Gender Matter? Decision Making in State Supreme Courts. *Social Science Quarterly*, *81*(3), 750–762.

Supreme Court of the United States. (n.d.). *Visitor's guide to oral argument*. https://www.supremecourt.gov/visiting/visitorsguidetooralargument.aspx

Thölke, P., Mantilla-Ramos, Y.-J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., Kemtur, A., Mekki Berrada, L., Sahraoui, M., Young, T., Bellemare Pépin, A., El Khantour, C., Landry, M., Hadid, V., O'Byrne, J., Jerbi, K., Pascarella, A., & Combrisson, E. (2023). Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *NeuroImage*, *277*. https://doi.org/10.1016/j.neuroimage.2023.120253

Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-10358-x