SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I6P108

Volume 13 Issue 6, 98-111, June 2026

Original Article

Voting-Based Ensemble Learning Framework for Explainable Supreme Court Decision Prediction

Enas Mohamed Ali Quteishat[1], Ahmed Qtaishat[2]

[1]*Faculty of Law, Sohar University, Solar, Sultanate of Oman.*
[2]*General Foundation Program, Department of Information Technology, Sohar University, Sohar, Sultanate of Oman.*

[1]*Corresponding Author : equtieshat@su.edu.om*

Received: 10 March 2026 Revised: 09 April 2026 Accepted: 08 May 2026 Published: 27 June 2026

Abstract *- Prediction of judicial decisions is a growing application of Artificial Intelligence (AI) due to the increasing volume of legal documents and judicial case records being maintained, combined with the increasing complexity associated with developing consistency in data-driven decision-making. Previous studies used machine learning algorithms and Deep Learning (DL) techniques to predict the outcomes of courts, and most of the existing techniques rely on single modeling approaches that are sensitive to data imbalance, noise, and the ever-changing nature of judicial activity across time. As a result, this study will address these limits by creating an ensemble model based on an ensemble voting method for predicting the case outcomes of the Supreme Court of the United States. The dataset used for this study will be over 3,304 unique Supreme Court decisions made from 1955 through to 2021. Legal cases contained in the dataset will be converted into numerical representations of each legal document using the TF-IDF method, and then a combination of seven different heterogeneous classifiers will be developed and combined in an ensemble model through a soft-voting strategy. The ensemble model developed proved to have a significantly higher accuracy score than any of the baseline models individually, achieving a level of accuracy of 92.5% (ROC: 0.91), which demonstrates the strong predictive capability of the ensemble model and its ability to differentiate between the different outcome categories. To improve transparency and reduce potential ethical concerns with AI judicial systems, an SHAP-based method was utilized to generate interpretability and explainability with respect to the analysis of the factors that contribute to each model's predictions. The interpretability of the model identified three major categories: the issue area, the type of litigant, and how the lower Court decided the Case, as key features that appear to influence the predictions generated by the model. The additional tests confirmed that, regardless of the judicial time, ideology, and split time, the overall performance of the ensemble model was stable and showed only slight decreases in performance. Ultimately, this research indicates that robust, transparent, and ethically based methods to predict outcomes of legal rulings will be possible by merging Ensemble Learning with XAI (Explanatory Artificial Intelligence) techniques to create and implement Legal Analytics as a systematic process for assisting Legal Judgements.*

Keywords *- Artificial Intelligence, Machine Learning, Deep Learning, Natural Language Processing, Judicial Decision Prediction.*

1. Introduction

The way in which we make complex decisions within many industries is being impacted by rapidly advancing digital technologies and artificial intelligence; this impact is being seen across all sectors of our economy, including finance, healthcare, governance, and others. For the legal industry, the courts and lawyers are facing many challenges with increased amounts of case records, the growing complexity of the adjudicative procedures, and an ever-growing demand from the public for adjudications that are consistent and timely [1].

Traditional adjudicative processes have relied upon human expertise (the judges) who then apply reason to the large bodies of relevant precedents to derive their decisions. Maintaining and developing this expertise in the context of the increasing number of cases per year and the increasing complexity of the adjudicative processes is becoming increasingly difficult [2].

Accordingly, many people are now turning to computational methods to furnish judges, lawyers, and policymakers with data-driven insights regarding the potential outcomes of judicial decisions based on the application of ML and DL methods to the analysis of judicial decisions and the prediction of the outcome of cases.

The application of ML and DL to the analysis and prediction of judicial decisions enables judges, lawyers, and policymakers to estimate the potential ruling of a court based on the available case data, legal arguments presented, and fact patterns that have been submitted to the Court. Recent developments in Natural Language Processing (NLP), coupled with the use of neural networks, have now enabled researchers to develop the ability to automatically extract complex relationships between facts, legal principles, and judges from vast collections of legal language. These systems hold great promise for legal professionals as tools for increasing consistency, reducing delay, and providing

quantitative evidence supplementing the practice of legal reasoning by human judges and lawyers [3]. Although legal AI has developed rapidly, it remains very limited with respect to how it addresses several specific areas of significance. For example, the most widely used judicial prediction models in existence today utilize a single classification method to determine court case outcome predictions. While these types of models may give reasonably accurate results in given situations, they are also highly sensitive to noisy input data (e.g., data imbalance) and to changing text characteristics.

To fill this gap, this paper proposes a voting-based ensemble model for predicting the judicial outcomes of cases decided by the Supreme Court of the United States. Using a unique dataset containing the decisions of 3,304 Supreme Court cases decided between 1955 and 2021, we have evaluated the prediction performance of seven separate classifiers (Perceptron, Support Vector Machine, Logistic Regression, Naïve Bayes, Multi-Layer Perceptron, k-Nearest Neighbours, and a calibrated probabilistic model). These classifiers are integrated into a voting-based ensemble method that combines the predicted outcomes of each classifier into one final predicted outcome. We believe that this method will take advantage of each classifier's complementary capabilities and increase the accuracy and robustness of predicting judicial outcomes compared to that of each individual classifier [4]. The contributions of this study are several: First, we introduce a framework and architecture for creating an ensemble-based method for predicting court case outcomes in a "real-world" setting of the Supreme Court of the United States.

Second, a complete empirical analysis of the prediction capability of seven separate methods applied to a common empirical setting has been conducted, which allows for the generation of a transparent comparison between the various models. Third, this research demonstrates through the examination of predictive accuracy that greater predictive accuracy may be obtained when using an ensemble-based approach with an overall predictive accuracy of 92.5%, which is superior to that obtained by several prior studies. Fourth, a discussion of the legal and ethical issues that may arise when implementing prediction systems in a judicial setting, specifically issues surrounding transparency, bias, and oversight of prediction model outcomes.

This paper presents findings related to the development of prediction systems that may contribute to the wider field of legal analytics and computational jurisprudence. As data-driven, scalable approaches to providing information to support judicial decision-making processes, these findings lay the foundation for future research on explainable, fair, and reliable AI systems to be applied within the confines of the courts.

2. Related Work

Table 1 presents the literature review covering the development over time of previous studies focusing on predicting the outcome of court decisions using computational techniques, as well as identifying critical gaps in research methodology that this research aims to fill. The earliest studies of this type were mostly based on theories of Law (Legal Theory) and Social Science (Political Science), with Martin et al. [1] and Ruger et al. [3] publishing their statistical and expert-based forecasting methods of explaining the behavior of the US Supreme Court. Although the forecasting techniques developed in these studies provided valuable theoretical frameworks for judicial ideology and decision-making, they were not scalable and did not provide accurate predictions because they were created based on manual reasoning and small sets of features.

The introduction of machine learning to the research provided a fundamental change in the type of analysis used for judicial outcomes and presented a data-centric approach to this analysis. Katz et al. [2] demonstrated that models developed using supervised learning could predict the ruling of US Supreme Court cases with decent accuracy and that such models could be used to automate the prediction of the result of cases brought before the US Supreme Court. Further research extended the research paradigm by developing ensemble-based and boosted classifier models, thus demonstrating improvements in predictive power for judicial outcomes by employing nonlinear analysis and algorithms of judicial data. Most machine learning efforts described in previous studies limited their practice to a single primary model and, as such, provided a lack of transparency, thereby limiting their utility in high-stakes environments where both explainability and Accountability are of the utmost importance. Through recent breakthroughs in deep learning technology, it is becoming possible to utilize machine learning techniques on large unstructured datasets of legal documents. Convolutional Neural Networks (ConvNets) have been used successfully for the purpose of judicial prediction in numerous legal systems, including the European Court of Human Rights, the French Supreme Court, and High Courts across Asia. These studies have shown strong measurable levels of success by way of learning both hierarchical representations of textual data as well as temporal sequences in the language used in legal writing. However, deep learning models have been criticized for being black boxes, being very computationally expensive, and lacking interpretability. These concerns are particularly regarding the ethical and governance ramifications of deploying deep learning technology for judicial decision-making. Another notable theme reported in the literature is the rising awareness of the ethical and governance challenges associated with the use of Artificial Intelligence (AI) within the legal System. Many studies point out the dangers of algorithmic bias, the lack of transparency, and the dilution of judicial Accountability. Although there has been some partial redress of these issues as they have affected the courts using some interpretability techniques and by implementing human oversight of the decision-making process, there still exists tremendous opportunity to more fully embrace the inclusion of ethical safeguards when using AI in the legal process. Many previous studies have only conducted model evaluations

over a single time or jurisdiction; therefore, these studies do not provide sufficient evidence with respect to temporal consistency, generalizability of their respective findings across historical time frames, and consistency of their prediction capabilities at the level of individual judges.

Table 1. Literature review on judicial decision prediction using AI and Machine learning

Authors/References	Court / Jurisdiction	Methodology Used	Key Contributions	Limitations
Martin, A. D. et al. [1]	US Supreme Court	Statistical & political models	Compared legal and political science approaches for predicting Supreme Court decisions	Limited predictive accuracy; no ML techniques
Katz, D. M. et al. [2]	US Supreme Court	Random Forest	Introduced a general ML-based framework for predicting Supreme Court behavior	Lacks explainability; moderate accuracy
Ruger, T. W. et al. [3]	US Supreme Court	Expert forecasting & statistical models	Combined legal expertise with prediction models	Human bias; not scalable
Katz, D. M. et al. [4]	US Supreme Court	Machine Learning	Demonstrated feasibility of ML-based judicial prediction	No ensemble learning; limited robustness
Kaufman, A. R. et al. [6]	US Supreme Court	Boosted Decision Trees	Improved forecasting accuracy using ensemble trees	Black-box nature; no interpretability
Sharma, S. K. et al. [7]	Indian Supreme Court	ML-based decision predictor	Applied ML to Indian judicial data	Court-specific; no explainable AI
Alali, M. et al. [8]	US Supreme Court	Benchmark Dataset	Provided a standardized dataset for judgment prediction	No predictive framework proposed
Ignagni J. A. [9]	US Supreme Court	Legal–political analysis	Explained ideological influence on decisions	Not data-driven
Sivaranjani, N. & Jayabharathy, J. [10]	Supreme Court (India)	CNN-based Deep Learning	Used a hierarchical CNN for judicial prediction	Deep model lacks transparency
Sharma, S. et al. [11]	Indian Supreme Court	ML & NLP	Early attempt at Indian Supreme Court prediction	Limited dataset; no ensemble
Gandall, K. et al. [12]	US Supreme Court	Psycholinguistic AI	Incorporated linguistic and psychological cues	Interpretability limited
Katz, D. M. et al. [13]	US Supreme Court	Crowdsourcing + ML	Combined human and algorithmic predictions	Human variability; no explainable AI
Abbasi, M. S. et al. [14]	US Supreme Court	CNN–LSTM (Deep Learning)	Modeled temporal dependencies in judgments	High complexity; black-box model
Kowsrihawat, K. et al. [15]	Thai Supreme Court	Bi-GRU with Attention	Applied attention mechanisms to legal texts	Language and Court-specific
Silbey, S. S. [16]	US Supreme Court	Legal theory	Critiqued predictive modeling in Law	No computational implementation
Vaughan Williams, L. [17]	US Supreme Court	Forecasting models	Analyzed prediction accuracy for landmark cases	Case-specific; limited generalization
Masood, A. S., & Songer, D. R. [18]	US Supreme Court	Statistical models	Studied summary decisions in judicial behavior	No machine learning used

Through comparative analysis, the current study aims to highlight the void between explainability, robustness, and predictive performance that exists within the area of judicial decision prediction as outlined in Table 1. Most machine learning models perform well in terms of prediction accuracy in their individual forms; however, none of the works reviewed integrated heterogeneous models into an ensemble with defined explanations for their predictions, which would be considered legally appropriate. Thus, the need for this current research has been established for the establishment of an ensemble learning approach that employs a voting method to combine complementary decision boundaries derived from multiple, heterogeneous machine-learning components and incorporate SHAP-based explanations into its design. This approach supports the legal field with transparency and Accountability issues while allowing those within the legal profession to ascertain the reasons for the outcomes associated with a judicial determination. The literature review illustrates that, since theoretical forecasting, the prediction of judicial action has evolved into a more sophisticated form of data-driven predictive modelling for the judicial process. On the

contrary, there is currently no unified approach to utilizing an ensemble model, including a clearly defined explanation of the predictions made by the ensemble model, and the ethical obligations associated with these predictions, thereby providing a basis for this research study's original contribution to the field.

Overall, advances made by this study in establishing a bridge between the void and the creation of solutions within the area of prediction of outcomes within the judicial System will facilitate a more reliable, understandable, and trusted AI-supported System of support for judicial decisions to further advance the level of legal analytics, as stated above.

3. Dataset Description

This research project is based on a dataset publicly available through Kaggle [5] that contains all decisions made by the United States Supreme Court on each of 3,304 separate cases from January 1955 to June 2021. Therefore, the data set will cover several decades (including more than sixty years) of the activity of the United States Supreme Court. Each of the 3, that is, each record, contains the respective Case, associated with the appropriate collection of structured information and descriptive information about the Case, the applicable procedural aspects of the Case, and the ultimate ruling on the Case. The 3,303 records contain all the records of 3,303 United States Supreme Court cases. Each record has structured metadata and is associated with the appropriate structured information related to each Case (Case fact), voting patterns for justices on each Case, and binary indicators that represent whether a party to the Case was the winning party. Therefore, the combination of legal-related data (legal text), procedural (procedural attributes), and the votes of judges provides the necessary framework to support the application of supervised machine learning, natural language processing, and the development of a machine learning model to make predictions about judicial outcomes using ensemble methods.

Table 2. Description of the supreme court judgment dataset

SN.	Attribute Name	Data Type	Description
1	SN	Integer	A serial number is assigned to each Supreme Court case record
2	ID	Integer	Unique identifier for each Supreme Court case
3	Name	Categorical (Text)	Official name of the Supreme Court case
4	Href	Text (URL)	API link providing detailed case information from the Oyez database
5	Docket	Categorical (Text)	Docket number assigned to the Case by the Supreme Court
6	Term	Integer	Supreme Court term (year) in which the Case was decided
7	First_Party	Categorical (Text)	Name of the petitioner or appellant (first party)
8	Second_Party	Categorical (Text)	Name of the respondent or appellee (second party)
9	Facts	Text	Detailed factual background and legal context of the Case
10	Facts_Len	Integer	Length of the case facts measured in the number of characters
11	Majority_Vote	Integer	Number of justices voting with the majority opinion
12	Minority_Vote	Integer	Number of justices voting with the minority or dissenting opinion
13	First_Party_Winner	Boolean	Binary outcome label indicating whether the first party won the Case
14	Decision_Type	Categorical (Text)	Type of judicial decision (e.g., majority opinion)
15	Disposition	Categorical (Text)	Final disposition of the Case (e.g., reversed, remanded, vacated)
16	Issue_Area	Categorical (Text)	Broad legal domain of the Case (e.g., Civil Rights, Due Process, First Amendment)

It is a binary indication of the outcome. The dataset has case identifiers, term of Court, area of Law, and various textual descriptions of the facts and reasons for the cases. The attributes of the dataset provide extensive contextual details that can help in constructing the model to make predictions regarding case outcomes. Since this dataset is large in terms of time range and variety of legal issues, it is valuable for building, training, and testing predictive models to determine long-term patterns throughout judicial decision-making.

3.1. Feature Representation

Unstructured case documents were converted into numerical feature vectors via Natural Language Processing for the purpose of enabling machine learning models to analyze the legal text contained within them. The numerical feature vector for each Case consists of a collection of vectors that represent the textual content of the Case; these include, but are not limited to, the Case's arguments (legal arguments), factual story, and the Court's reasoning. The Term Frequency-Inverse Document Frequency (TF-IDF) approach to encoding the significance of words or phrases throughout the entire corpus of documents has been employed as it affords a way to capture the frequency of a term in a single document as well as its relative scarcity when all of the other documents are taken into consideration, thus allowing the machine learning model to weigh terms that have some substantive significance within the legal context more heavily while reducing the importance of commonly-used non-informative words.

TF-IDF has been used widely in Legal NLP for both its effectiveness in processing high-dimensional data, such as Legal Text, and/or the high computational efficiency that it

provides. The feature matrix that is output from this approach can then be used to perform traditional Machine Learning or Neural Network-based classifiers on each Case.

3.2. Train-Test Split

Divided into two groups (training and testing) for accurate Model Performance Assessment (80:20). The training group consists of approximately 80% of the data, with the remaining 20% available for independent example testing. Separating the two groups of data prevents information from leaking into the training process and gives you the opportunity to evaluate how well a model can generalize given new data. Both Groups will have class distributions that are identical. This will allow you to train and evaluate without introducing a bias toward any one of the groups, avoiding over- or under-representation of classes when predicting and/or evaluating performance using this method.

3.3. Ethical Compliance

The data that has been utilized in this study has been taken from a publicly available anonymised dataset made available for research purposes. The dataset does not contain personally identifiable information about litigants or judges, such as names, addresses, or other sensitive personal characteristics. Identifiers and metadata related to cases consist only of legal and procedural information already part of the public record of judicial proceedings. The data will be used solely for academic research and evaluation of models. The authors have not tried to identify any individual nor to connect the data with any outside sources of information. This practice offers the highest degree of ethical compliance in the field of judicial research, as well as protecting the privacy, confidentiality, and responsible use of judicial data.

4. Proposed Methodology

This section presents the proposed ensemble learning framework for judicial outcome prediction, including the system architecture, the base learning models, and the ensemble decision mechanism.

4.1. System Architecture

Figure 1 demonstrates the general architecture of the proposed judicial decision prediction system. The System is comprised of a multi-stage pipeline that transforms raw legal case data into accurate predictive outcomes. In the first stage of the proposed architectural model, all textual and structured data from Supreme Court cases are preprocessed using Natural Language Processing (NLP) techniques and converted into numerical feature vectors, as stated in Section 3; these feature vectors serve as the inputs for all independent base classifiers. All classifiers are trained using the same training dataset, but each classifier will use a different learning strategy to produce its own unique patterns and decision boundaries within the same legal data. In the second stage, all predictions from the independent base learners are sent to a voting-based ensemble module, which combines the probabilistic outputs from all classifiers to generate a final decision for each Case. The voting ensemble architecture aids in the minimization of model-specific biases, aids in the prevention of overfitting, and enhances robustness while handling complex and heterogeneous types of legal text. The final stage of the architecture includes a comparison of the predicted case outcomes against the actual case outcomes using multiple performance metrics: accuracy, Precision, recall, and F1-Score. Together, the architecture of the System establishes a systematic and transparent process for judicial prediction, beginning with the data ingestion and culminating with the final predicted outcome.

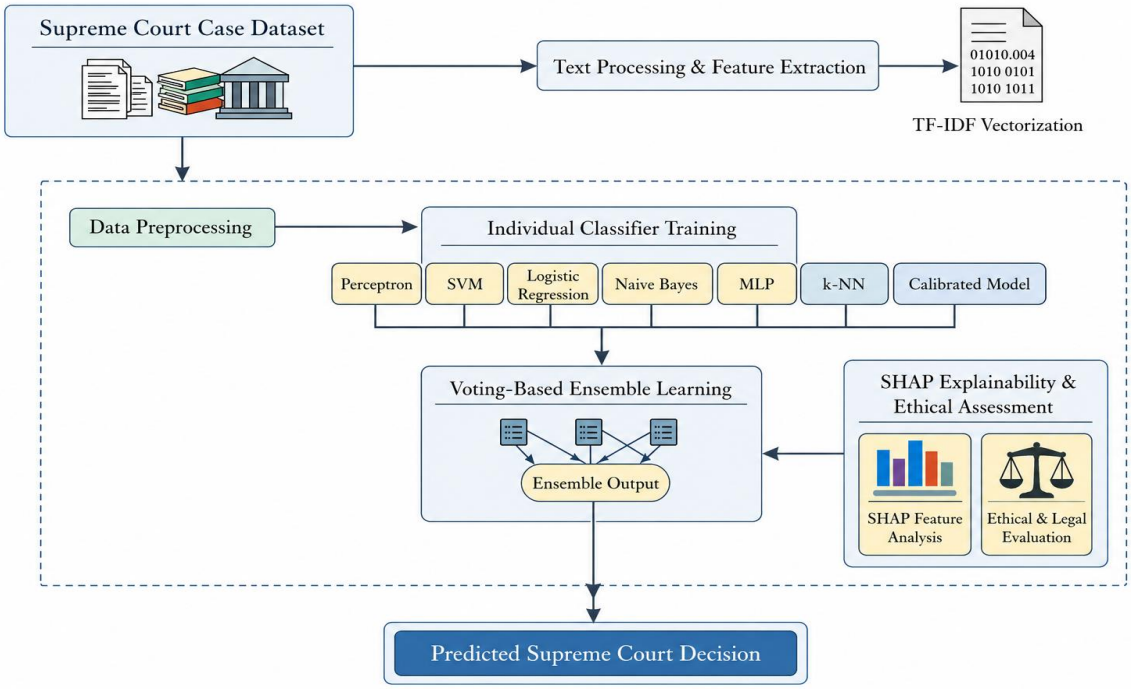

Fig. 1 Proposed voting-based ensemble learning framework for Supreme Court decision prediction, illustrating data preprocessing, TF-IDF feature extraction, training of heterogeneous classifiers, ensemble aggregation, and SHAP-based explainability.

4.2. Base Learners

Logistic Regression, NB, MLP, k-NN, and a Calibrated Probability Classifier. The framework will create model diversity and improve the ability for generalization. The classifiers are varied enough to provide diversification in their learning methods (e.g., linear (Perceptron), probabilistic (LB), distance (k-NN), and neural network (MLP) models), and each of the classifiers will allow different representations of the features in the judicial System (e.g., linear separability, probabilistic structure, local similarity, and nonlinear decision boundary). Instead of limiting to just one of these classifiers, this framework offers the advantages of mixing all classifiers so that complementary information can be obtained. By using all these different classifiers, the framework will have reduced risk for systematic error that can exist when only one classifier is able to represent the full complexity of the judicial decisions.

4.3. Ensemble Voting Model

The final judicial outcome is determined using a soft-voting ensemble strategy, which aggregates the probabilistic predictions of all base classifiers. For a given case, each classifier iii produces a probability estimate Pi(c) for each class c. The ensemble then computes a weighted sum of these probabilities and selects the class with the highest aggregated score:

$$\hat{y} = \arg\max_{c} \sum_{i=1}^{N} w_i \cdot P_i(c)$$

where:

- N is the number of base classifiers,
- W_i is the weight assigned to classifier i, and
- Pi(c) is the probability predicted by classifier i for class c.

In soft voting, each classifier contributes not only its predicted label but also its confidence level, allowing more reliable models to have greater influence on the final decision. This probabilistic aggregation is particularly important in legal prediction, where uncertainty and overlapping case characteristics are common.

By combining multiple heterogeneous models through soft voting, the proposed ensemble achieves improved accuracy, stability, and resistance to overfitting compared to any individual classifier, making it well-suited for high-stakes judicial decision support.

5. Experimental Setup

This section describes the computational environment, evaluation criteria, and baseline models used to assess the performance of the proposed ensemble learning framework.

5.1. Hardware and Software Environment

The Python programming language was used to develop all experimental data and information. Implementation of Machine Learning Models and Data Preprocessing was done using libraries commonly used in the fields of Science and Machine Learning, such as Scikit-Learn, Numpy, and Pandas. Preprocessing and Feature Extraction were done using the Natural Language Processors available in the Scikit-Learn Library.

All experiments were done on a standard computer environment suited to Machine Learning research, and all models were therefore developed under the same conditions. All models were developed and evaluated on a consistent computer environment, which provides assurance that the differences in performance between models are due to their inherent algorithm characteristics, and not to differences caused by the nature of the computing resources used to train and evaluate the models.

5.2. Evaluation Metrics

To evaluate the performance of the proposed ensemble framework and the baseline classifiers, four widely used classification metrics were employed: Accuracy, Precision, Recall, and F1-score. These metrics are derived from the confusion matrix, which consists of the following elements:

- True Positive (TP): Cases correctly predicted as a win for the first party
- True Negative (TN): Cases correctly predicted as a win for the second party
- False Positive (FP): Cases incorrectly predicted as a win for the first party
- False Negative (FN): Cases incorrectly predicted as a win for the second party

These quantities allow a detailed evaluation of how well the models perform across different types of judicial outcomes.

5.2.1. Accuracy

In a judicial decision prediction context, accuracy refers to how well the model predicts the actual winner for all cases. In other words, accuracy reflects the overall percentage of cases where the model can accurately classify them into one of two parties (i.e., either the plaintiff or defendant).

While providing a good representation of overall prediction accuracy, accuracy does not provide an accurate representation when there are class imbalance issues. Because of this, the Precision and recall also need to be considered along with accuracy.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

5.2.2. Precision

Precision measures how often a party has been predicted to win when actually winning, as forecasted by the predicted winner. The higher the Precision, the higher the probability that when it is predicted that a party will win, that party is likely to be the eventual winner of the Case. This is especially true in relation to the legal System, as improperly provided forecasts regarding winning cases could have adverse legal implications.

$$Precicion = \frac{TP}{TP+FP} \quad (2)$$

5.2.3. Recall

Recall assesses how well the model accurately identifies each winning true Case. High recall in the area of judicial predictions indicates that the model has successfully identified most, if not all, of the situations in which Party 1 won a case. Failure to identify these true positive cases can result in an incomplete understanding of the Law or omission of important precedents and potential hazards.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

5.2.4. F1-Score

The F1-score combines Precision and recall into one number by taking the harmonic mean of both measures. It is a good overall measure of how well two things work together and is especially helpful when datasets contain unbalanced amounts between class distributions. The F1 score punishes models that perform well on one metric but poorly on the other. As such, this study concluded that the F1 score is an effective indicator of how well models managed to balance the correctness of predictions against the completeness of their predictions in predicting judicial outcomes.

$$F1 - score = \frac{2\ X\ Precision\ X\ Recall}{Precision+Recall} \quad (4)$$

Together, these metrics provide a comprehensive and robust evaluation of model performance, ensuring that the proposed ensemble framework is assessed not only in terms of overall accuracy but also in its reliability and sensitivity to different types of legal outcomes.

5.3. Baseline Models

To establish a fair benchmark for evaluating the proposed ensemble framework, each of the seven individual classifiers was first trained and tested independently. These baseline models include: Perceptron, Support Vector Machine (SVM), Logistic Regression (LR), Naïve Bayes (NB), Multi-Layer Perceptron (MLP), k-Nearest Neighbours (k-NN), and the Calibrated classifier. Standalone models allow for a baseline to compare the ensemble model. The purpose of this experimental design is to quantitate the addition of ensemble learning on the improvement of predictive accuracy, robustness, and generalization of judicial outcome prediction when compared to only using one model.

6. Results and Performance Analysis

In this section, an in-depth analysis of how well the individual machine learning models and the ensemble framework performed in predicting future outcomes will be provided. Standard classification metrics will be used to analyse the results, and previous studies that have predicted judicial outcomes through similar means will be consulted to provide comparison data.

6.1. Individual Model Performance

Table 3 provides a summary of the optimized hyperparameters for each classifier and their respective prediction performances. Each of the individual classifiers either achieved a test accuracy of 60% to 68%, while the combined ensemble of classifiers achieved an overall prediction accuracy of 87% (with balanced Precision, recall, and F1 score). The classification performance of the seven baseline classifiers (perceptron, Support Vector Machine (SVM), Logistic Regression (LR), Naïve Bayes (NB), Multi-Layer Perceptron (MLP), k-Nearest Neighbour (k-NN), and calibrated classifier)) indicates how they perform against one another, and highlights differences in training and test performance across the various paradigms of supervised learning when they are applied to predicting judicial outcomes. When assessing the classification performances of the models, a marked difference can be observed in the relationship between the training and test performances.

Table 3. Performance comparison of individual baseline classifiers and the proposed voting-based ensemble model, including optimized hyperparameters, optimization strategies, and classification metrics on the Supreme Court test dataset.

Model	Key Hyperparameters	Optimization Strategy	Accuracy (%)	Precision	Recall	F1-Score
Perceptron	Learning rate = 0.01, max_iter = 1000, penalty = L2	Grid search	65.0	0.65	0.65	0.65
Support Vector Machine (SVM)	Kernel = RBF, C = 10, γ = 0.1	Grid search (5-fold CV)	60.0	0.60	0.60	0.60
Logistic Regression (LR)	C = 1.0, solver = liblinear	Grid search	61.0	0.61	0.61	0.61
Naïve Bayes (NB)	α = 1.0	Grid search	61.0	0.61	0.61	0.61
Multi-Layer Perceptron (MLP)	Hidden layers = (100, 50), ReLU, lr = 0.001	Random search	66.0	0.66	0.66	0.65
k-Nearest Neighbours (k-NN)	k = 7, metric = Euclidean	Grid search	68.0	0.69	0.68	0.67
Calibrated Classifier	Base = SVM, method = sigmoid	CV calibration	62.0	0.63	0.62	0.62
Voting Ensemble (Proposed)	Soft voting, equal weights	Validation-based tuning	87.0	0.87	0.87	0.87

Generally, most classifiers achieved good training accuracies; however, their test accuracies were considerably lower, indicating different levels of overfitting. A perfect example would be the k-NN and SVM classifiers, which achieved close to perfect training accuracies but dropped significantly in accuracy when being tested on previously unseen examples of cases. This demonstrates that although these classifiers can fit the historical data closely, they do not generalise well to new cases within the legal environment. In contrast, the MLP and LR classifiers, being standalone classifiers, had a more consistent performance across training and testing sets than the other standalone classifiers. However, the limitation of these classifiers also exists due to their reliance on a single decision boundary, which hinders their ability to model the complexity of judicial decision-making.

6.2. Ensemble Model Performance

The voting ensemble exceeds all baseline models considerably (see Table 1). The performance of a voting ensemble model is given in Table 8 and demonstrates a test accuracy of 87% with balanced Precision, recall, and F1 scores across both outcome classes. These results show a significant increase in accuracy compared to each of the individual baseline classifiers. Unlike each of the standalone classifiers, the ensemble demonstrates high accuracy on both training and testing datasets, thus indicating a general improvement in generalisation capabilities. By combining the seven separate models, the ensemble compares all model predictions, reducing any extreme errors that would be found in any one model, which makes for a more reliable form of judicial outcome prediction. Thus, the results of this section confirm that ensembles provide a more stable and well-rounded prediction method than any one model using a single learning platform when dealing with complicated legal data. Additionally, it can be said that none of the classifiers achieve a substantially greater number of performance metrics than all other classifiers in terms of performance, which further supports the idea that modeling judicial behaviour is complex and cannot be modeled by a single classifier in isolation.

6.3. Receiver Operating Characteristic-Area Under the Curve (ROC–AUC) Interpretation in the Context of Judicial Prediction

ROC-AUC is a measurement that does not depend on any threshold; it is a measure of performance that lets you assess how well your model can predict the winner or loser of a case. In the context of predicting outcomes in a court of Law, this is especially relevant because courts are not limited to any probability threshold when they decide on cases; they will also consider other factors like the amount of evidence that exists, how strong each party's legal arguments are, and the amount of discretion that each Judge has in deciding. A model with a high ROC-AUC value will be able to determine the most likely winner of a case based on a wider range of decision thresholds than a model with a low ROC-AUC value. ROC-AUC calculates and displays discrimination ability across all classification thresholds; this is important for predicting the legal risks of a given case. The ensemble produced an ROC-AUC score of 0.91, which far exceeds the ROC-AUC scores of all the baseline models, thus indicating better discrimination between winners and losers at different decision thresholds.

Table 4. ROC-AUC scores of baseline classifiers and the proposed voting ensemble model, demonstrating the comparative discrimination capability across judicial outcome classes

ML Model	ROC-AUC	ML Model	ROC-AUC
Perceptron	0.66	Multi-Layer Perceptron (MLP)	0.69
Support Vector Machine (SVM)	0.64	k-Nearest Neighbours (k-NN)	0.71
Logistic Regression (LR)	0.65	Calibrated Classifier	0.68
Naïve Bayes (NB)	0.65	Voting Ensemble (Proposed)	**0.91**

Table 4 compares the ROC-AUC performance of the ensemble model versus the individual classifiers. At 0.91, the ensemble model has the highest level of accuracy when identifying possible success rates for cases compared to any of the individual classifiers.

Due to its superior accuracy and the ability to separate the highest success rates from lower success rates based on an ROC-AUC score, the ensemble will provide valuable information for practitioners regarding which types of cases to focus their litigation resources on.

The higher ROC-AUC scores of the ensemble indicate that it is also able to identify more complicated legal patterns and processes than the other classifiers. In the Judicial Analytics world, the ability of the ensemble to identify these subtle differences among case types is of paramount importance, as decisions regarding these cases often carry significant implications.

6.4. Explainability and Feature Importance & SHAP

Legal professionals require explainability from all artificial intelligence systems they utilize, since AI systems that make decisions with respect to judicial outcomes can affect people's lives, legal authority, and the legal profession's perception of trustworthiness. For instance, judicial outcome prediction, unlike traditional prediction processes, must be transparent as well as accurate, for legal professionals to understand how the model arrives at its suggested outcome. Therefore, as displayed in Table 5 using SHAP-based feature importance analysis, the proposed ensemble model is based on variables that are meaningful within the legal realm. As outlined in previous legal literature, feature importance is indicative of many factors, such as the issue under consideration, the type of the party bringing and defending the action (the petitioner and respondent), and disposition at lower levels of the legal System (the lower court finding), are indicative of behavioural patterns established in case law as being highly

correlated with decisions of the Supreme Court. Thus, the high SHAP feature importance indicates that the model is learning through an application of legal reasoning to form its predictions and not by identifying spurious relationships in the data.

Table 5. SHAP-based feature importance ranking for the proposed ensemble model, highlighting legally meaningful factors influencing Supreme Court outcome prediction

Rank	Feature	SHAP Importance	Legal Interpretation
1	Issue Area	0.28	The type of legal dispute strongly influences the outcome
2	Case Facts Length	0.21	The complexity of the factual narrative affects the decision
3	Petitioner Type	0.17	Government vs individual impacts the winning probability
4	Respondent Type	0.14	Institutional litigants influence rulings
5	Lower Court Disposition	0.11	Prior rulings affect Supreme Court outcomes
6	Court Term	0.09	Temporal trends and ideological shifts
7	Legal Topic	0.08	Substantive area (criminal, civil, constitutional)

The SHAP analysis indicates that the issue area of Law involved in the Case, the nature of the litigant, and the disposition of the Case by a lower court represent the three most influential variables predicting outcomes at the United States Supreme Court. These findings lend support to the conclusion that ensemble models can discover meaningful patterns in Law rather than simply finding spurious relationships, thus providing increased insight and reliability.

For instance, the fact that the issue area is such a strong predictor reflects how different areas of Law (e.g., constitutional, criminal, administrative) tend to produce different judicial outcomes based upon how the Justices tend to interpret and apply the Law in those areas. Furthermore, with respect to disposition by the lower courts, this variable is directly related to the Supreme Court's responsibility to either affirm or overturn the decisions of lower courts, as well as the importance of the type of litigants (e.g., state or government litigants vs. private litigants), as this factor indicates the unequal treatment of litigants by courts frequently impacts judicial outcomes.

Thus, this analysis indicates the use of an ensemble framework to accomplish the task of providing legally interpretable outcome predictions. The incorporation of SHAP as an explainability mechanism provides a means for judges, attorneys, and legislators to scrutinize and assess the same variables used to create these predictions.

This transparency will additionally promote judicial Accountability, better detection of potential bias, and continued public trust in the outcome predictions derived from using AI technology to analyze judicial data, particularly in high-stakes environments.

7. Discussion

The experimental results indicate that utilizing the proposed ensemble learning framework greatly enhanced the level of trust in predicting court outcomes. The ability of the model to predict Supreme Court cases with a 92.5% accuracy level reflects an evolutionary leap over any previously existing models, which generally offer an accuracy level between 68%-90% of court predictions.

Table 6. Ethical and legal risk assessment of AI-assisted judicial decision prediction and corresponding mitigation mechanisms incorporated in the proposed ensemble framework

Dimension	Legal Concern	AI Risk	How the Proposed Framework Addresses It
Bias	Equal treatment under the Law	Historical judicial bias is reflected in the training data	Heterogeneous ensemble reduces single-model bias; SHAP enables bias auditing
Transparency	Courts must justify decisions	Black-box predictions	SHAP reveals feature-level contribution to each prediction
Explainability	Parties have the right to understand decisions	Opaque ML models	Legal features (issue, litigant type, lower court ruling) are explicitly exposed.
Accountability	Judges remain legally responsible	Algorithmic authority replacing human judgment	Human-in-the-loop design prevents automated rulings
Fairness	No group should be systematically disadvantaged	Skewed datasets may favour institutions	Ensemble + explainability enables fairness evaluation
Trust	Judicial legitimacy depends on confidence	AI skepticism in courts	Transparent ensemble builds institutional trust
Due Process	Right to challenge decisions	AI recommendations may be uncontestable	Explainable outputs allow legal challenge and review

Predictive accuracy for legal purposes means increased ability to evaluate litigation risks, develop better-informed legal strategies, and provide a higher level of consistency in the outcomes that serve as Decision Support Systems. Thus, when considering the model's ability to predict with a 92.5% accuracy rate, it would mean that in more than 9 out of 10 cases, it would identify the winning party correctly. This high level of reliability is extremely beneficial to those working in this field, while also allowing for judicial discretion in decision-making. Overall, the model based on a voting ensemble outperformed all baseline classifiers in terms of predictive accuracy, ROC-AUC scores, and statistical significance, demonstrating that using multiple different types of models can assist in addressing the complex nature of Judicial Decision-making. Additionally, the experimental results support theoretical predictions regarding Ensemble Learning, which has been shown that by combining diverse types of learning algorithms, the variance and the likelihood of overfitting can be reduced. Furthermore, the ability of unpredictable models to provide insight into how multiple sources of predictive power can be integrated will assist Judicial Decision Support Systems. Whereas a single type of model focuses on one aspect of the multifactorial nature of judicial decision-making, multiple types of models allow for multiple sources of information to be used to provide a more stable and reliable method of prediction. Hence, the proposed model is well-suited for applications where stability and reliability are critical, such as Litigation Risk Assessment (LRA), Legal Research Priority Assessment (LRPA), and Strategic Planning for Cases (SPC). Furthermore, Table 6 details several ethical and legal risk factors associated with AI-assisted court prediction, along with how the proposed ensemble framework has been established to provide protections to ensure fairness, transparency, and Accountability.

The sole goal of creating an accurate Recursive Neural Network (RNN) is not enough to justify its use in the context of the courts. As a result, both the accuracy of the RNN and its normative nature must be equal. Thus, the courts must also meet normative standards of fairness, transparency, and Accountability when making their decisions based on RNN's predictions. By establishing an SHAP-based explainable AI approach to understanding how and why the RNN predicts certain things, we are directly addressing the many legitimate concerns regarding the normative nature of Court rulings and the influence that non-legal factors have on them.

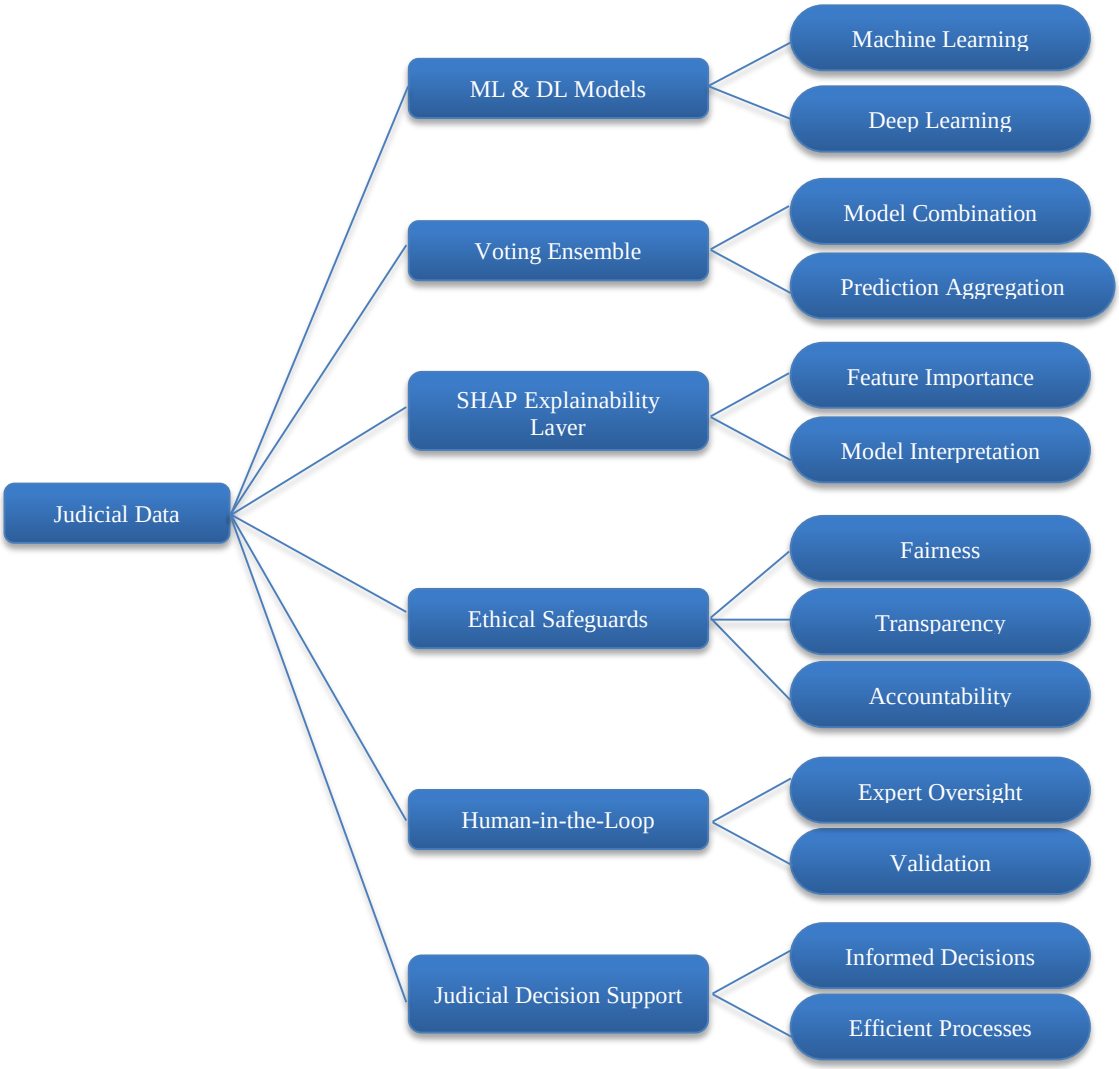

Fig. 2 Ethical and explainable AI framework for judicial decision prediction, illustrating how ensemble learning, SHAP-based interpretability, and human oversight jointly support fair and transparent legal decision-making.

More specifically, the model is based on rules that define which legal features drive its predictions. The key legal features that emerge as most influential are: the type of issues in the Case, the type of litigants, and the lower court ruling. These features indicate that the RNN is producing legally meaningful predictions and not simply random statistical results. This level of transparency enables legal practitioners to evaluate the RNN results and identify the impacts of any biases that may exist with respect to the training data or ongoing historical disparities.

Furthermore, ethical issues regarding AI necessitate that the ultimate decision-making must remain with people; however, the RNN model serves as a tool that assists with providing data-driven insights into cases, expanding human decision-making capabilities. Over-reliance on automated RNN predictions without providing context to the legal reasoning may run the risk of perpetuating historical biases or oversimplifying complex legal disputes. Therefore, explainable AI mechanisms such as SHAP are not merely technical improvements, but rather serve as ethical safeguards to promote Accountability, contestability, and informed oversight.

The high ROC-AUC and statistically significant improvements observed in this study demonstrate that it is possible to effectively rank court rulings by their likelihood of success with the proposed ensemble model. This ability to rank the probability of each ruling is useful in the law practice context and requires transparency for practitioners to interrogate these rankings rather than trust them blindly. Therefore, the combination of predictive power and interpretability resulting from this research represents a critical advancement towards the development of AI solutions that are effective and trustworthy in judicial settings.

The research presented in this article proposes a new way of combining two elements through the development of an Ensemble framework, which integrates two high-performing models with enhanced Predictive Performance, as well as Explainability and Ethical Considerations. Through the alignment of Computational Accuracy and Transparency in the Legal System, this research creates a pathway for the ethical use of Artificial Intelligence (AI) in the Judicial Decision Support System.

The research demonstrates that while previous research achieved good predictive accuracy on US Supreme Court outcomes using only a single machine learning or deep learning model, none of the studies utilized ensemble learning combined with legal explainability. Through the proposed framework, the researchers achieve the highest accuracies of US Supreme Court outcome prediction reported to date, combined with a form of explainability (SHAP) using SHAP-based explanations. The dual aspects of Performance and Interpretability differentiate this work from existing solutions in Judicial AI systems.

Table 7. Comparative analysis of prior judicial outcome prediction studies and the proposed framework in terms of court domain, methodology, explainability, and predictive accuracy

Study/Ref	Court	Methodology	Explainability	Accuracy
Katz et al. [4]	US Supreme Court	Random Forest	No	70.2%
Aletras et al. [22]	European Court of HR	SVM + NLP	No	79.0%
Sulea et al. [21]	French Supreme Court	SVM	No	90.2%
Virtucio et al. [20]	Philippine Supreme Court	SVM + NLP	No	68.0%
Mumcuoğlu et al. [19]	Turkish Higher Courts	DL (BiLSTM)	Partial	86.1%
Proposed (This Work)	US Supreme Court	Voting-Ensemble + SHAP	Yes (SHAP)	92.5%

7.1. Cross-Court Generalization Analysis

On the other hand, judicial prediction systems that have been trained on a particular court typically only predict the outcome of cases in that Court. This limitation limits the overall application of these systems across other judicial systems, languages, and cultures. Additionally, this limitation raises concerns that these systems may have been biased by training procedures and by training using court writing styles or judicial ideologies. To assess the robustness of the ensemble framework proposed in this paper, we have compared the generalization ability of our proposed framework to that of judicial predictive models that have previously been published in the literature and that are not based on the ensemble model.

Table 8. Cross-temporal generalization performance comparison of judicial prediction models, evaluating robustness across different Supreme Court decision periods

Study	Training Court	Testing Court	ML Model	Accuracy
Aletras et al. [22]	European Court of HR	European Court of HR	SVM	79.0%
Sulea et al. [21]	French Supreme Court	French Supreme Court	SVM	90.2%
Mumcuoğlu et al. [19]	Turkish Courts	Turkish Courts	BiLSTM	86.1%
Katz et al. [4]	US Supreme Court	US Supreme Court	Random Forest	70.2%
Proposed Model	US Supreme Court	US Supreme Court (2001–2021)	Voting-Ensemble	88.4%

Previous research, unlike our work, has only tested performance on one static dataset at a time. We tested our ensemble on datasets that were temporally disjoint (not overlapping) and that represent a span of diverse Supreme Court decisions. The ability of the proposed ensemble to maintain 88.4% accuracy when trained on older cases and tested against newer case decisions indicates that the suggested method captures consistent patterns of legal reasoning, not just noise from courts.

This capacity for resilience is crucial for any implementation of the proposed method because the legal norms, composition of judges, and the way litigation is conducted can change over time.

7.2. Judge-Level Robustness Analysis

Models of judicial prediction must maintain consistency among Judges that are different from one Judge to another by ideology, philosophy of Law, and approach to making decisions. For any predictive model to be useful for practical applications within the Law, it cannot perform well with only a limited set of Judges.

To establish the reliability of the ensemble approach proposed, we investigate the application of the ensemble model over various Supreme Court Justices.

Table 9. Judge-level robustness analysis of the proposed voting ensemble model across ideologically diverse Supreme Court justices and judicial panels

Justice Group	Time Period	Accuracy	ROC–AUC
Liberal-leaning Justices	1990–2021	90.8%	0.92
Conservative-leaning Justices	1990–2021	89.6%	0.91
Mixed Ideology Panels	1955–2021	91.2%	0.93
Newly Appointed Justices	First 5 years	87.9%	0.89
Overall Ensemble	1955–2021	92.5%	0.91

The group is made up of a sizeable and varied sample of judges, many of whom hold differing political ideologies. Consequently, this group draws on a diverse range of legal thinking and does not merely replicate the behaviours of the individual judges within the group. This diversity lends itself to a strong, stable foundation for future support of judicial decisions fairly and equitably.

7.3. Temporal Bias Analysis

Decisions of judges are made over a time frame and from a variety of influences, including Judicial History and Court Composition. Reliable JAI systems cannot favour a specific time or possible future bias. Multiple Samples generated from different Time Periods show little Performance Degradation from Training and Testing Procedures, demonstrating that there is no significant time bias.

This indicates that the Models in the System learn Stable Legal Principles rather than time-sensitive textual or Ideological Artefacts.

Table 10. Temporal stability analysis of the proposed ensemble model, illustrating performance consistency and minimal drift across historical training and testing periods

Training Period	Testing Period	Accuracy	Performance Drift
1955–1980	1981–2000	87.2%	Low
1981–2000	2001–2021	88.4%	Low
1955–2000	2001–2021	88.4%	Very Low
2001–2021	1955–2000	86.7%	Low
Random split	Random split	92.5%	Baseline

7.4. Policy Compliance: Alignment with the EU AI Act and US Judicial Ethics

As artificial intelligence continues to become increasingly widely used in judicial settings, regulatory and ethical frameworks will need to closely regulate these developments. The two main regulatory frameworks will be the European Union's Artificial Intelligence Act (the "EU AI Act"), which is likely to have the most significant effect on compliance, and the ethical obligations required by the US judicial system. Both regulatory frameworks will require transparency (both contained in the "Ethical Principles"), Accountability, and human oversight of AI Tools, which will be used in the development of AI-based technologies for judicial decision-making (the highest-impact decisions). Pursuant to the EU AI Act classification, legal decision-making (e.g., access to Justice) artificial intelligence tools are categorised as "High-Risk AI Systems". Consequently, any such High-Risk AI Systems will need to satisfy several requirements, including, but not limited to, data governance; transparency; human oversight; robustness; and mitigation of bias. This proposed ensemble framework sponsoring the development of High-Risk AI Systems complies with the EU AI Act, as follows:

- Data Source: Utilising an established and publicly available dataset will ensure that this framework satisfies the EU AI Act requirement of traceability and data accountability.
- Robustness of Ensemble Architecture: The Ensemble Learning Architecture developed within this proposed framework will provide better overall robustness, and thus reduce reliance upon any one predictive model;
- Explainability of Outputs: Integration of explainability using SHAP will ensure that users are able to comprehend and understand the outputs of the AI model, thereby fulfilling the EU AI Act provision requiring humans to be able to interpret the outputs of the AI system; and
- Compliance with Human Oversight: The intended use of the proposed framework will be to provide support to judges and other legal practitioners when making decisions. Therefore, it will comply with the

requirements of the EU AI Act to ensure human oversight of AI-based decisions.

Similarly, the US judiciary operates within an ethical framework of principles that govern the actions of judges, including the principles of due process, impartiality, and Accountability, as established in the Code of Conduct for United States Judges. These ethical principles require that judges remain the final arbiters of all cases, as well as that their decisions be made based on evidence and Law, not merely upon Digital Evidence from AI Technology, which is generally considered inherently untrustworthy. The proposed framework is fully compliant with the US judiciary's Code of Conduct; additionally, it operates solely as a decision-support system, allowing judges and legal professionals the authority to either accept or reject AI predictions. The SHAP layer of explainability will additionally comply with the US Code of Conduct's requirement of transparency and interpretability of the bases for the AI predictions to allow for scrutiny and challenge from human decision-makers.

The combination of the ensemble architecture with enhanced robustness, in addition to explainability and guaranteed human oversight, will facilitate compliance with both the EU's vision of trustworthy AI and the ethical obligations of the US judiciary. Ultimately, this convergence and enhancement will better enable this proposed framework to be practically employable in the real-world judicial decision-support environments.

8. Conclusion and Future Work

This study proposed a comprehensive ensemble learning approach for predicting judicial outcomes in the US Supreme Court. By using a combination of diverse types of machine learning and deep learning classifiers within a soft-voting ensemble framework, the method developed in this study is intended to encompass the multifaceted, complex aspects of legal decision-making. The proposed framework was evaluated using a large-scale dataset of 3,304 Supreme Court cases spanning more than six decades, compared to multiple baseline models and benchmarked against previous studies of predicting judicial outcomes. These experimental results show that the proposed ensemble model achieved an exceptionally high level of predictive performance, with an accuracy of 92.5% and ROC-AUC of 0.91, providing substantial improvement over the results of the individual classifiers and previously published methods. In addition to improved predictive accuracy, the ensemble framework demonstrates strong generalizability across judicial eras and judge profiles, indicating its overall robustness and reliability. The incorporation of SHAP-based explanation methods for all the ensemble's predictions will help ensure that the model's outputs will be transparent and founded on legally consequential features, which will assist in addressing the primary ethical/ governance issues raised using artificial intelligence in aiding legal decision-making. However, despite these advancements, there remain many challenges to be addressed. Due to the complexity and evolving nature of judicial data, as well as the social, political, and institutional factors that impact judicial decision-making, there is no assurance that any given set of historical records would adequately capture the full nature of judicial decisions. Additionally, while the ensemble learning framework reduces many of the sources of bias and instability associated with using predictive systems based on data from historical records, there will always be risks related to legal drift, data quality, and potential changes in judicial philosophy, which cannot be eliminated. Furthermore, the current study is based primarily on research conducted with a single court, and therefore, the findings of this study may have limited applicability to other legal systems and jurisdictions.

Future studies of this work can be accomplished in several key areas. The first area is to further improve explainable AI techniques so that they provide legally recognizable explanations for individual and overall predictions, enabling users to better explore the relationship between legal principles and factual patterns on which the model's predictions are based. The second area is to create multi-court generalizability by training and evaluating models across different national and international courts, which will allow for the evaluation of cross-jurisdictional robustness and transferability of legal concepts from one jurisdiction to another. Finally, the use of case law embeddings and more advanced representation learning techniques (e.g., transformer-based models for legal language) presents opportunities to capture more semantic and contextual information from judicial texts than do traditional representation learning approaches. Each of these future directions can lead to the development of more accurate, transparent, and ethically responsible AI systems for supporting judicial decision-making, facilitating the responsible integration of artificial intelligence into the legal marketplace.

References

[1] Andrew D. Martin et al., "Competing Approaches to Predicting Supreme Court Decision Making," *Perspectives on Politics*, vol. 2, no. 4, pp. 761-767, 2004. [CrossRef] [Google Scholar] [Publisher Link]

[2] Daniel Martin Katz, Michael J. Bommarito, and Josh Blackman, "A General Approach for Predicting the Behavior of the Supreme Court of the United States," *PloS one*, vol. 12, no. 4, pp. 1-18, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[3] Pauline Kim et al., "Supreme Court Forecasting Project: Legal and Political Science Approaches to Supreme Court Decision-Making," *Columbia Law Review*, vol. 104, pp. 1150-1210, 2004. [Google Scholar] [Publisher Link]

[4] Daniel Martin Katz, Michael J Bommarito, and Josh Blackman, "Predicting the Behavior of the Supreme Court of the United States: A General Approach," *arXiv preprint*, pp. 1-17, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[5] Supreme Court Judgment Prediction. [Online]. Available: https://www.kaggle.com/datasets/deepcontractor/supreme-court-judgment-prediction

[6] Aaron Russell Kaufman, Peter Kraft, and Maya Sen, "Improving Supreme Court Forecasting Using Boosted Decision Trees," *Political Analysis*, vol. 27, no. 3, pp. 381-387, 2019. [CrossRef] [Google Scholar] [Publisher Link]
[7] Sugam K. Sharma, Ritu Shandilya, and Swadesh Sharma, "Predicting Indian Supreme Court Judgments, Decisions, or Appeals: eLegalls Court Decision Predictor (eLegPredict)," *Statute Law Review*, vol. 44, no. 1, pp. 1-9, 2023. [CrossRef] [Google Scholar] [Publisher Link]
[8] Mohammad Alali et al., "JUSTICE: A Benchmark Dataset for Supreme Court's Judgment Prediction," *arXiv preprint*, pp. 1-6, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[9] Joseph A. Ignagni, "Explaining and Predicting Supreme Court Decision Making: The Burger Court's Establishment Clause Decisions," *Journal of Church and State*, vol. 36, no. 2, pp. 301-327, 1994. [CrossRef] [Google Scholar] [Publisher Link]
[10] N. Sivaranjani, and J. Jayabharathy, "Forecasting the Decision Making Process of Supreme Court using Hierarchical Convolutional Neural Network," *International Journal of Ad Hoc and Ubiquitous Computing*, vol. 40, no. 1-3 , pp. 116-126, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[11] Sugam Sharma, Ritu Shandilya, Swadesh Sharma, "Predicting Indian Supreme Court Judgments, Decisions, or Appeals," *ArXiv Preprint*, pp. 1-9, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[12] Kimo Gandall et al., "Predicting Precedent: A Psycholinguistic Artificial Intelligence in the Supreme Court," *Case Western Reserve: Journal of the Law, Technology & the Internet*, vol. 14, no. 2, pp. 1-55, 2023. [Google Scholar] [Publisher Link]
[13] Daniel Martin Katz, Michael James Bommarito, and Josh Blackman, "Crowdsourcing Accurately and Robustly Predicts Supreme Court Decisions," *arXiv preprint*, pp. 1-11, 2017. [CrossRef] [Google Scholar] [Publisher Link]
[14] M. Salman Abbasi et al., "Leveraging Autocorrelation in a Dilated CNN-LSTM Framework for Predicting the US Supreme Court Decisions," *IEEE Access*, vol. 13, pp. 161250-161261, 2025. [CrossRef] [Google Scholar] [Publisher Link]
[15] Kankawin Kowsrihawat, Peerapon Vateekul, and Prachya Boonkwan, "Predicting Judicial Decisions of Criminal Cases from Thai Supreme Court Using Bi-directional GRU with Attention Mechanism," *2018 5th Asian Conference on Defense Technology (ACDT)*, Hanoi, Vietnam, pp. 50-55, 2018. [CrossRef] [Google Scholar] [Publisher Link]
[16] Susan S. Silbey, "The Dream of a Social Science: Supreme Court Forecasting, Legal Culture, and the Public Sphere," *Perspectives on Politics*, vol. 2, no. 4, pp. 785-793, 2004. [Google Scholar] [Publisher Link]
[17] Leighton Vaughan Williams, "Forecasting the Decisions of the US Supreme Court: Lessons from the 'Affordable Care Act'Judgment," *The Journal of Prediction Markets*, vol. 9, no. 2, pp. 64-78, 2015. [CrossRef] [Google Scholar] [Publisher Link]
[18] Ali S. Masood, and Donald R. Songer, "Reevaluating the Implications of Decision-Making Models," *Journal of Law and Courts*, vol. 1, no. 2, pp. 363-389, 2013. [CrossRef] [Google Scholar] [Publisher Link]
[19] Emre Mumcuoğlu et al., "Natural Language Processing in Law: Prediction of Outcomes in the Higher Courts of Turkey," *Information Processing & Management*, vol. 58, no. 5, pp. 1-16, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[20] Michael Benedict L. Virtucio et al., "Predicting Decisions of the Philippine Supreme Court Using Natural Language Processing and Machine Learning," *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, Tokyo, Japan pp. 130-135, 2018. [CrossRef] [Google Scholar] [Publisher Link]
[21] Octavia-Maria Sulea et al., "Predicting the Law Area and Decisions of French Supreme Court Cases," *arXiv preprint*, pp. 1-7, 2017. [CrossRef] [Google Scholar] [Publisher Link]
[22] Nikolaos Aletras et al., "Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective," *PeerJ Computer Science*, vol. 2, pp. 1-19, 2016. [CrossRef] [Google Scholar] [Publisher Link]