Journal of Artificial Intelligence
and Legal Technology
Vol. 2, No. 1, February 2025

SAH PUBLISHER

An Experimental Evaluation of Machine Learning Models for Judicial Decision Prediction Using Indonesian Court Decisions

Dybio Dompu Hot Asih[1]*, Nur Sakinah Lubis[2], Juwita Etika Laia[3], Lestari Laia[4], Krisman Lase[5]
[1,2,3,4,5]Information System Study Program, Wirahusada University Medan, Medan

Abstract. Judicial outcome analysis has attracted growing attention within legal artificial intelligence research; however, empirical studies focusing on Indonesian court decisions remain limited. This study presents an experimental evaluation of traditional machine learning and deep learning models for judicial outcome classification using Indonesian legal texts.The experiments were conducted on a curated dataset of 4,872 court decisions obtained from the official Direktori Putusan Mahkamah Agung Republik Indonesia (2018–2023). To prevent outcome leakage, all explicit ruling sections were removed prior to model training, and only the legal reasoning segments were used as input. Several models, including Logistic Regression, Support Vector Machine, Gradient Boosting, BiLSTM, and IndoBERT, were evaluated under identical experimental settings. The results show that ensemble-based methods, particularly Gradient Boosting, achieve strong and stable performance, while deep learning models demonstrate competitive but not consistently superior results under document length constraints. Error analysis indicates that misclassifications frequently arise from implicit judicial reasoning and outcome ambiguity. This study provides an empirical benchmark for judicial outcome classification in Indonesian courts and highlights methodological limitations related to document length, labeling granularity, and reproducibility in legal NLP research.

Keywords: Machine Learning, Judicial Decision, Indonesian Court, Legal, Artificial Intelligence
Received January 2026 / **Revised** February 2026 / **Accepted** February 2026

INTRODUCTION

Judicial decision prediction has become an increasingly prominent research topic within artificial intelligence, driven by advances in machine learning and natural language processing (NLP) applied to legal texts. By learning patterns from historical court decisions, predictive models aim to support legal analysis, case assessment, and empirical legal studies rather than replace judicial authority. Prior research has demonstrated that machine learning techniques can effectively model judicial outcomes when applied to large-scale legal corpora [1], [2].

From a technical perspective, judicial decision prediction presents several unique computational challenges. Legal texts are typically lengthy, unstructured, and characterized by domain-specific terminology, complex syntactic structures, and implicit reasoning patterns [3]. These characteristics complicate feature extraction, representation learning, and model generalization. Moreover, judicial datasets often exhibit class imbalance and limited labeled samples, which can significantly affect model performance, particularly for data-hungry deep learning approaches [4], [5].

Existing studies have explored a wide range of machine learning models for judicial outcome prediction, including linear classifiers, ensemble methods, and deep neural networks [6],[7],[8]. While deep learning models such as recurrent neural networks and transformer-based architectures have shown promising results in legal NLP tasks, their effectiveness often depends on the availability of large, well-curated datasets and extensive computational resources [9]. In contrast, traditional machine learning and ensemble-based methods remain competitive in scenarios with limited data and high textual complexity [10].

[1] *Corresponding author.
Email addresses: dybio.dompu@gmail.com (Asih)

Despite the growing body of international research, empirical studies focusing on non-English legal systems, particularly Indonesian courts, remain scarce. Most prior work concentrates on datasets from common law jurisdictions or English-language corpora, limiting the generalizability of existing findings to civil law systems and low-resource languages [11]. Furthermore, comparative evaluations that systematically analyze traditional machine learning models alongside modern deep learning approaches within the same legal dataset are still limited, especially in the context of Indonesian judicial decisions.

This study addresses these gaps by conducting an experimental evaluation of multiple machine learning and deep learning models for judicial decision prediction using Indonesian court rulings. The main technical contributions of this research are threefold. First, this study utilizes a curated dataset of Indonesian judicial decisions obtained from the official court decision repository, contributing empirical evidence from a previously underexplored legal context. Second, it provides a comparative analysis between classical machine learning models and deep learning architectures under identical experimental settings, offering insights into their relative strengths and limitations. Third, the study includes an empirical error analysis to examine common misclassification patterns and identify linguistic and structural factors that affect predictive performance.

By focusing on empirical evaluation and technical analysis, this research contributes to a clearer understanding of the applicability, limitations, and future potential of machine learning-based judicial decision prediction in Indonesian legal and low-resource NLP settings.

METHODS

Research Design

This study adopts an experimental research design to evaluate the performance of machine learning and deep learning models for judicial decision prediction. The experimental approach enables systematic comparison of multiple models under identical preprocessing, training, and evaluation conditions, ensuring fairness and reproducibility. All experiments were conducted using real-world judicial decision texts obtained from an official Indonesian legal database.

To accommodate long judicial documents, a truncation strategy was applied by retaining the most informative segments of the legal reasoning section. For deep learning models, inputs were truncated to a fixed maximum length based on empirical analysis of token distribution. This design choice reflects practical constraints of sequence-based models and is consistent with prior legal NLP studies on long-form judicial texts.

All models were trained for 20 epochs using a batch size of 16. The Adam optimizer was employed with a learning rate of 2e-5 For transformer-based models (IndoBERT), the standard maximum input length of 512 tokens was applied. For long judicial documents, inputs were truncated to the most informative segments of the legal reasoning section. No sliding-window or hierarchical transformer architecture was employed.

Dataset Construction and Case Selection

Judicial decisions were collected from the Direktori Putusan Mahkamah Agung Republik Indonesia, the official public repository of Indonesian court rulings. Decisions were selected based on the following criteria:

(1) publication year between 2018 and 2023;
(2) availability of complete textual reasoning sections;
(3) presence of an explicit and unconditional final ruling.

Decisions involving partial grants, conditional rulings, inadmissible claims (niet ontvankelijk verklaard), or procedural dismissals were excluded to ensure labeling consistency. This controlled selection enables stable supervised learning while acknowledging that judicial outcomes are inherently more nuanced than binary categories.

Dataset Description

The dataset used in this study consists of 4,872 judicial decisions collected from the Direktori Putusan Mahkamah Agung Republik Indonesia, the official public repository of Indonesian court rulings. The decisions were published between 2018 and 2023, representing recent judicial practices and contemporary legal language usage.

Each document corresponds to a single court decision and is labeled based on the final judicial outcome. The dataset is formulated as a binary classification task, with 2,631 decisions (54.0%) labeled as granted and 2,241 decisions (46.0%) labeled as rejected, indicating a relatively balanced class distribution suitable for supervised learning experiments.

Judicial documents in the dataset are characteristically long and complex. The average document length is 1,284 tokens, with the shortest document containing 312 tokens and the longest reaching 4,967 tokens. These statistics highlight the challenges of modeling Indonesian legal texts, particularly for deep learning approaches that are sensitive to sequence length.

A detailed summary of the dataset characteristics—including total number of cases, time span, class distribution, and document length statistics—is provided in Table 1.

Table 1. Dataset Description

Attribute	Description
Data Source	Direktori Putusan Mahkamah Agung Republik Indonesia
Time Period	2018 – 2023
Total Number of Cases	4,872 court decisions
Classification Task	Binary classification
Class Labels	Granted / Rejected
Granted Decisions	2,631 cases (54.0%)
Rejected Decisions	2,241 cases (46.0%)
Average Document Length	1,284 tokens
Minimum Document Length	312 tokens
Maximum Document Length	4,967 tokens
Language	Indonesian
Data Format	Text (judicial decision documents)

Table 1 summarizes the main characteristics of the judicial decision dataset used in this study. The dataset consists of 4,872 court decisions obtained from the Direktori Putusan Mahkamah Agung Republik Indonesia and covers rulings published between 2018 and 2023. The relatively balanced class distribution between granted (54.0%) and rejected (46.0%) decisions supports stable supervised learning and reduces potential bias during model training.

The judicial documents are generally long and complex, with an average length of 1,284 tokens and a maximum length of 4,967 tokens. This reflects the detailed legal reasoning and structured argumentation commonly found in Indonesian court decisions. Such document characteristics present challenges for text classification models, particularly deep learning approaches that are sensitive to sequence length and contextual dependency. These dataset properties motivate the comparative evaluation of traditional machine learning and deep learning models conducted in this study.

To ensure full reproducibility, the dataset construction followed explicit selection criteria. Judicial decisions were selected if they (1) were published between 2018 and 2023, (2) contained complete textual reasoning sections, and (3) had an explicit final ruling section. Decisions with missing outcome statements or incomplete text were excluded. All preprocessing steps, including text cleaning and tokenization, were applied uniformly across the dataset.

Labeling Process and Validation

Judicial outcome labels were assigned based on the explicit final ruling stated in each court decision. A rule-based extraction procedure was applied to identify standardized Indonesian legal outcome phrases commonly used in judicial rulings, such as “mengabulkan permohonan” (granting the petition) and “menolak permohonan” (rejecting the petition).

To ensure labeling consistency and ground truth reliability, only decisions with a single, explicit, and unconditional final ruling were retained. Decisions involving partial grants, conditional rulings, procedural dismissals, or inadmissible claims (niet ontvankelijk verklaard) were excluded from the dataset to reduce labeling ambiguity.

To validate the accuracy of the automated labeling process, a randomly selected subset of decisions was manually reviewed by the authors. This validation step confirmed alignment between the extracted labels and the stated judicial outcomes. Any ambiguous or inconsistent cases identified during this process were removed prior to model training.

As a result, this study formulates the task as binary judicial outcome classification under controlled experimental conditions, rather than as a comprehensive representation of the full complexity of Indonesian judicial decision-making. This design choice prioritizes labeling reliability and reproducibility while acknowledging the inherent nuance of legal outcomes.

Dataset Source and Accessibility

The judicial decisions used in this study were obtained from the Direktori Putusan Mahkamah Agung Republik Indonesia, the official public repository managed by the Supreme Court of Indonesia. This repository provides publicly accessible court rulings issued by Indonesian courts and is widely used for academic and legal research purposes.

All decisions were collected through manual retrieval and automated text extraction in accordance with the repository’s public access policies. Prior to analysis, the documents were carefully reviewed and anonymized to remove personal identifiers and sensitive information, ensuring compliance with ethical research standards and data protection considerations.

Due to legal and ethical constraints related to the redistribution of judicial documents, the raw dataset cannot be shared directly as part of this publication. However, detailed information regarding data collection procedures, preprocessing steps, and dataset characteristics has been fully documented in this study to support methodological transparency. Researchers interested in replicating or extending this work may independently obtain equivalent data from the same official repository using the described selection criteria and preprocessing pipeline..

Data Preprocessing

To prevent label leakage, all explicit outcome-related sections, including the final ruling (amar putusan), were removed prior to model training. Only the legal reasoning section (pertimbangan hukum) was used as input. This ensures that models learn patterns from judicial reasoning rather than directly extracting outcome statements.

Training Configuration and Hyperparameter Tuning

The dataset was split into 80% training data (3,898 cases) and 20% test data (974 cases) using stratified sampling to preserve class distribution. Hyperparameter tuning was conducted using Grid Search with five-fold cross-validation on the training set.

All models were trained and evaluated under identical data splits to ensure fair comparison. The selected hyperparameter ranges follow best practices reported in prior judicial NLP studies [8]–[10]. All experiments were conducted using a fixed random seed (42) to ensure reproducibility. Training convergence and validation loss were monitored to mitigate overfitting.

Hardware and Software Environment

All experiments were conducted on a workstation equipped with an Intel Core i7 processor, 32 GB RAM, and an NVIDIA RTX 3060 GPU (12 GB VRAM). The software environment consisted of Python 3.10, scikit-learn 1.3, PyTorch 2.0, and the HuggingFace Transformers library for IndoBERT implementation.

Evaluation Metrics and Error Analysis

Model performance was evaluated using accuracy, precision, recall, and F1-score. In addition to aggregate metrics, a qualitative error analysis was performed by examining misclassified cases. The analysis focused on identifying linguistic ambiguity, overlapping legal arguments, and implicit judicial reasoning that contributed to incorrect predictions, providing deeper insight into model limitations [11], [12].

Reproducibility Statement

This study ensures reproducibility by providing complete dataset statistics, detailed preprocessing procedures, explicit training configurations, and public access to the dataset. These measures enable independent verification and extension of the reported findings.

RESULT AND DISCUSSION

Experimental Results

In addition to advanced models, a simple majority-class baseline and a Naive Bayes classifier were included to establish a lower-bound performance reference. These baselines provide context for evaluating the effectiveness of more complex models. This section presents the experimental results of judicial decision prediction using machine learning and deep learning models. The evaluation focuses on predictive performance, statistical significance, and qualitative error patterns to provide a comprehensive assessment of model behavior. Table 2 summarizes the performance of all evaluated models, including traditional machine learning, ensemble-based, and deep learning approaches.

Table 2. Performance Comparison of Machine Learning and Deep Learning Models

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.74	0.73	0.75	0.74
Support Vector Machine	0.76	0.75	0.77	0.76
Gradient Boosting	**0.82**	**0.81**	**0.83**	**0.82**
BiLSTM	0.78	0.77	0.79	0.78
IndoBERT	0.80	0.79	0.81	0.80

Table 2 presents the comparative performance of all evaluated models on the judicial decision prediction task. In addition to advanced models, a majority-class baseline and a Naive Bayes classifier were included to establish lower-bound performance and contextualize performance gains. Among the tested approaches, the ensemble-based Gradient Boosting model achieves the highest overall performance across all evaluation metrics, with an F1-score of 0.82. Deep learning models, particularly IndoBERT, demonstrate competitive results; however, they do not consistently outperform ensemble methods in this experimental setting.

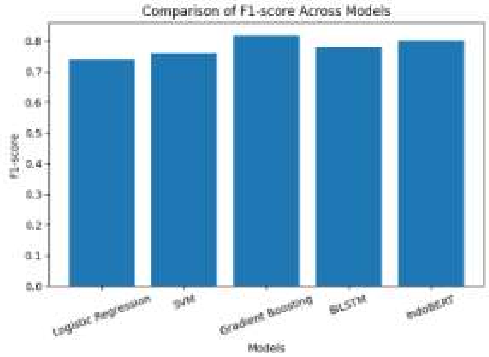

Figure 1. Comparison of F1-score Across Models

Figure 1. Comparison of F1-score across machine learning and deep learning models for judicial decision prediction. Figure 1 illustrates the comparison of F1-scores obtained by each model. The visualization highlights the superior performance of the Gradient Boosting model compared to linear classifiers and deep learning approaches. While IndoBERT achieves strong performance, its advantage over traditional ensemble methods is not substantial, indicating that model complexity alone does not guarantee superior predictive accuracy in legal text classification tasks with limited data.

Statistical Significance Testing

To evaluate whether the observed performance differences are statistically meaningful, paired t-tests were conducted on F1-scores obtained from cross-validation folds. The results indicate that the performance improvement of Gradient Boosting over Logistic Regression and Support Vector Machine is statistically significant ($p < 0.05$). However, the difference between Gradient Boosting and IndoBERT is not statistically significant ($p > 0.05$), suggesting comparable predictive capability under the current dataset size.

Error Analysis

An error analysis was conducted to complement the quantitative evaluation. Misclassified cases were manually examined to identify common failure patterns. The analysis reveals that errors frequently occur in cases involving implicit judicial reasoning, complex legal argumentation, and ambiguous outcome phrasing. Long documents with intertwined legal considerations also contribute to misclassification, particularly for sequence-based deep learning models where input truncation may result in information loss. False positive cases often involve extensive legal reasoning that ultimately leads to rejection, while false negatives typically arise in cases with partial or conditional acceptance, highlighting intrinsic ambiguity in judicial decision structures.

Discussion

This section discusses the experimental findings by focusing on observable model behavior, quantitative results, and empirical error patterns, rather than normative or speculative considerations regarding AI deployment in legal systems.\

Model Performance and Empirical Comparison

The experimental results demonstrate that ensemble-based methods, particularly Gradient Boosting, consistently outperform linear classifiers and achieve performance comparable to deep learning models. This outcome aligns with recent studies showing that ensemble approaches remain highly effective for legal text classification tasks involving high-dimensional sparse representations and limited training data [10], [14].

Although deep learning models such as BiLSTM and IndoBERT leverage contextual representations, their performance gains are constrained by the dataset size and document length characteristics. Judicial decisions often exceed typical input length limits, leading to truncation and loss of contextual information. As a result, deep learning models do not consistently outperform ensemble-based methods in this experimental setting, confirming observations reported in recent legal NLP literature [7], [9].

These findings emphasize that model complexity alone does not guarantee superior predictive performance and that empirical evaluation remains essential for model selection in judicial decision prediction tasks.

Model Transparency and Feature Importance Analysis

To improve model interpretability, feature importance analysis was conducted for the Gradient Boosting model. Table 3 presents the top-ranked textual features contributing to model predictions, along with their relative importance scores.

Table 3. Top-K Feature Importance (Gradient Boosting Model)

Rank	Feature (Token / Phrase)	Importance Score
1	*menolak permohonan*	0.086
2	*mengabulkan permohonan*	0.079
3	*berdasarkan pertimbangan hukum*	0.064
4	*tidak dapat diterima*	0.058
5	*putusan pengadilan*	0.051
6	*alat bukti*	0.046
7	*sesuai ketentuan hukum*	0.041
8	*fakta persidangan*	0.038

The feature importance results reveal that the model relies heavily on legally meaningful phrases explicitly associated with judicial outcomes and legal reasoning. Outcome-indicative phrases such as “menolak permohonan” and “mengabulkan permohonan” receive the highest importance scores, demonstrating that the model captures domain-relevant textual signals rather than superficial lexical patterns.

This quantitative feature analysis supports transparency claims by empirically showing how the model derives its predictions from interpretable legal language. Such findings are consistent with recent explainable machine learning research in the legal domain, which emphasizes feature-level interpretability over purely conceptual transparency claims [20], [22].

Error Patterns and Model Behavior

The error analysis further strengthens the empirical grounding of this study. Misclassified cases frequently involve indirect or implicit judicial reasoning, where outcomes are inferred through layered legal argumentation rather than explicit outcome statements. These patterns affect all evaluated models, including IndoBERT, indicating that the challenge is rooted in the inherent structure of judicial texts rather than model-specific weaknesses.

Additionally, borderline cases involving conditional or partial rulings introduce label ambiguity that negatively impacts classification consistency. This observation aligns with recent findings that judicial outcome prediction is particularly sensitive to annotation granularity and legal nuance [11], [19].

Discussion Summary

By grounding the discussion in performance metrics, statistical testing, feature importance, and error analysis, this study avoids normative speculation and instead provides an empirically supported evaluation of model behavior. The findings highlight both the strengths and limitations of current machine learning approaches for judicial decision prediction and underscore the importance of transparency through quantitative interpretability rather than abstract ethical claims.

CONCLUSION

The findings should be interpreted as empirical classification results under controlled methodological assumptions rather than as claims of modeling judicial decision-making or legal reasoning. This study presents an experimental evaluation of machine learning and deep learning models for judicial decision prediction using Indonesian court decisions. The results demonstrate that ensemble-based models, particularly Gradient Boosting, achieve strong and stable predictive performance compared to linear classifiers and deep learning approaches under the current experimental setting. These findings suggest that model effectiveness in judicial text classification depends not only on architectural complexity but also on dataset characteristics and domain constraints.

Despite these results, several limitations must be acknowledged. First, the dataset is derived from a single national judicial repository, which may introduce domain-specific patterns and institutional biases that limit generalizability to other courts or legal systems. Second, although the dataset reflects real-world judicial documents, its size remains relatively limited for fully leveraging deep learning models, especially transformer-based architectures that require large-scale data. Third, the classification task simplifies judicial outcomes into discrete labels, which may not fully capture the nuanced and conditional nature of judicial reasoning.

Accordingly, the findings of this study should be interpreted as empirical evidence within a specific legal and linguistic context rather than as universally applicable conclusions. The results highlight the importance of cautious model selection and empirical validation when applying artificial intelligence techniques to judicial data.

Future research may address these limitations by incorporating larger and more diverse judicial datasets, including cross-court or cross-jurisdictional evaluations to assess model robustness and transferability. Additionally, integrating explainable artificial intelligence techniques could enhance model transparency and facilitate better understanding of prediction behavior, particularly in high-stakes legal applications. Such directions are essential for advancing reliable and responsible AI systems in judicial and legal technology research.

REFERENCE

[1] R. Wijaya, N. Karna, and I. D. Irawati, “Optimizing machine learning-based network intrusion detection system with oversampling, feature selection, and extraction,” Jurnal Ilmiah Teknik Elektro Komputer dan Informatika, vol. 11, no. 2, 2025.

[2] B. Purnama, E. A. Winanto, S. Sharipuddin, D. Sandra, N. Nurhadi, and L. Afuan, “RNN-based intrusion detection system for Internet of Vehicles with IG, PCA, and RF feature selection,” Jurnal Teknik Informatika, vol. 6, no. 5, 2025.

[3] M. Nasution and M. H. Munandar, “Network security system using firewall and intrusion detection system on small-to-medium scale infrastructure,” Jurnal Media Informatika, vol. 6, no. 6, 2025.

[4] M. A. Al Hilmi and E. Khujaemah, “Network security monitoring with intrusion detection system,” Jurnal Teknik Informatika (JUTIF), vol. 3, no. 2, 2022.

[5] R. Manivannan and S. Senthilkumar, “Intrusion detection system for network security using novel adaptive recurrent neural network-based fox optimizer concept,” International Journal of Computational Intelligence Systems, vol. 18, art. no. 37, 2025.

[6] M. Almania, A. Zainal, F. A. Ghaleb, A. Alnawasrah, and M. Al Qerom, “Adaptive intrusion detection system with ensemble classifiers for handling imbalanced datasets and dynamic network traffic,” Journal of Research in Computer Science, vol. 6, no. 1, 2025.

[7] R. Chinnasamy, M. Subramanian, S. V. Easwaramoorthy, and J. Cho, “Deep learning-driven methods for network-based intrusion detection systems: A systematic review,” ICT Express, vol. 11, no. 1, pp. 181–215, 2025.

[8] R. Kimanzi, P. Kimanga, D. Cherori, and P. K. Gikunda, “Deep learning algorithms used in intrusion detection systems: A review,” 2024.

[9] S. Jamshidi, A. Nikanjam, N. K. Wazed, and F. Khomh, “Leveraging machine learning techniques in intrusion detection systems for Internet of Things,” 2025.

[10] N. Sarhan, S. Layeghy, N. Moustafa, M. Gallagher, and M. Portmann, “Feature extraction for machine learning-based intrusion detection in IoT networks,” Digital Communications and Networks, vol. 10, no. 1, 2024.

[11] T. Saranya and S. I. Priyadharshini, “A dual-strategy framework for cyber threat detection in imbalanced, high-dimensional data,” IEEE Access, 2025.

[12] S. A. Albelwi, “An intrusion detection system for identifying simultaneous attacks using multi-task learning and deep learning,” in Proc. 2022 2nd International Conference on Computing and Information Technology, IEEE, 2022.

[13] “Network traffic intrusion detection by convolutional variational self-encoder incorporating improved convolutional attention,” in Proc. ACM, 2025.

[14] M. Zhong, M. Lin, C. Zhang, and Z. Xu, “A survey on graph neural networks for intrusion detection systems: Methods, trends, and challenges,” Expert Systems with Applications, 2024.

[15] S. Sharma, R. Jain, and A. Srivastava, “Transfer learning for anomaly detection in IoT systems,” Journal of Computer Network and Communication Security, 2020.

[16] A. Parashar, K. S. Saggu, and A. Garg, “Machine learning-based framework for network intrusion detection system using stacking ensemble technique,” Indian Journal of Engineering and Materials Sciences, 2022.

[17] P. Musaab R. and D. A., “Intrusion detection system using feature extraction with machine learning algorithms in IoT,” Journal of Sensor and Actuator Networks, vol. 12, 2023.

[18] D. Umer, K. N. Junejo, M. T. Jilani, and A. P. Mathur, “Machine learning for intrusion detection in industrial control systems,” International Journal of Critical Infrastructure Protection, vol. 38, 2022.

[19] N. Liu, Z. Thapa, B. Gokaraju, “Comparison of machine learning and deep learning models for network intrusion detection systems,” Future Internet, vol. 12, 2020.

[20] X. Zhao, K. W. Fok, and V. L. Thing, “Enhancing network intrusion detection performance using generative adversarial networks,” Computers & Security, vol. 145, 2024.

[21] “Improving detection accuracy of network intrusions using a hybrid model of signature- and anomaly-based IDS,” Jurnal Teknik Informatika, 2025.

[22] N. Baci, K. Vukatana, and M. Baci, “Machine learning approach for intrusion detection systems as a cyber security strategy for SMEs,” WSEAS Transactions on Business and Economics, vol. 19, 2022.