Scientific Reports

https://doi.org/10.1038/s41598-026-47149-7

Article in Press

Multimodal deep learning for international investment arbitration outcome prediction and bilateral investment agreement negotiation strategy optimization

Received: 2 August 2025

Accepted: 30 March 2026

Published online: 03 April 2026

Cite this article as: Wu H. & Xu J. Multimodal deep learning for international investment arbitration outcome prediction and bilateral investment agreement negotiation strategy optimization. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-47149-7

Hao Wu & Jiajun Xu

Multimodal Deep Learning for International Investment Arbitration Outcome Prediction and Bilateral Investment Agreement Negotiation Strategy Optimization

Hao Wu[1,a*], Jiajun Xu[1,b]

1 School of Public Administration, Hohai University (Jiangning Campus), Nanjing 211100, Jiangsu, China

Email addresses:

a asd18103929689@163.com

b 19839219999@163.com

***Corresponding author:**

Hao Wu (asd18103929689@163.com)

ABSTRACT

International investment arbitration has expanded at a remarkable pace over the past two decades, generating pressing demand for robust outcome prediction tools that can guide strategic decisions. This study presents a multimodal deep learning framework that fuses textual, numerical, and visual data to predict arbitration outcomes in the investor-state dispute settlement context. Our attention-based fusion architecture channels legal documents, macroeconomic indicators, and visual evidence through dedicated encoders capable of capturing intricate cross-modal dependencies that shape tribunal reasoning. Evaluated on 1,247 arbitration cases drawn from major international institutions, the multimodal model attains an overall accuracy of 86.7%, surpassing single-modality counterparts by 7.8 percentage points and conventional machine learning baselines by 14.6 percentage points. Feature importance analysis reveals that the quality of legal argumentation, dispute monetary value, and arbitrator panel composition rank among the most decisive determinants of outcomes. Beyond their technical value, these findings equip investors, host states, and legal counsel with evidence-based tools for strategic planning, while simultaneously foregrounding normative questions about fairness, transparency, and equitable access to predictive technologies in dispute resolution.

KEYWORDS

Multimodal Deep Learning; International Investment Arbitration; Outcome Prediction; Legal Analytics; Artificial Intelligence in Law; Predictive Analytics

I. Introduction

Over the past three decades, international investment arbitration has undergone a dramatic expansion that has reshaped the global landscape of economic dispute resolution [1]. Data compiled by the United Nations Conference on Trade and Development show that known treaty-based investor-state dispute settlement cases climbed from fewer than 50 in 2000 to more than 1,200 by 2023, a trajectory that underscores how central arbitration has become to the adjudication of cross-border investment conflicts [2]. This surge in caseloads poses real challenges for legal practitioners, policymakers, and scholars who need dependable forecasts of arbitration outcomes to inform their strategic choices.

Conventional approaches to predicting arbitration results have largely rested on qualitative legal reasoning or on quantitative models that struggle to capture the layered character of investment disputes [3]. Most prediction frameworks concentrate either on legal precedents or on statistical relationships, and neither pathway fully accommodates the interplay among textual legal arguments, numerical economic indicators, and visual evidence—all of which jointly shape how tribunals reach their conclusions [4]. These methodological gaps have opened the door for computational methods that can process richer, more heterogeneous bodies of information.

Multimodal deep learning architectures offer a promising path forward, one that moves beyond the constraints of existing analytical methods [5]. Where traditional machine learning pipelines handle a single data type at a time, multimodal frameworks can simultaneously ingest and integrate heterogeneous inputs—textual legal documents, numerical economic series, and visual evidence—to construct more holistic predictive models [6]. Breakthroughs in natural language processing, computer vision, and neural network design have already demonstrated impressive results in legal case outcome prediction across various jurisdictions [7], pointing toward substantial untapped potential within international arbitration.

What sets this study apart is the construction of a purpose-built multimodal deep learning framework tailored to international investment arbitration prediction, one that weaves together three distinct yet complementary data modalities. On the textual side, advanced natural language processing techniques parse arbitration documents—legal arguments, factual narratives, procedural submissions—to extract nuanced semantic features. In parallel, the numerical branch processes economic indicators, financial figures, and quantitative evidence that frequently weigh on tribunal deliberations. A visual processing module rounds out the architecture, analyzing charts, graphs, technical diagrams, and photographic materials that often prove pivotal in investment dispute proceedings.

The contributions of this work are twofold. On the technical front, we introduce what is, to our knowledge, the first multimodal deep learning architecture expressly designed for arbitration outcome prediction, extending transformer-based models to accommodate heterogeneous legal data. On the practical front, the framework

furnishes evidence-based tools through which legal practitioners, investors, and host states can evaluate case strengths and litigation risks with greater precision. At the same time, these predictive capabilities raise normative concerns—about equitable access to such technologies, the possible amplification of arbitrator biases, and implications for due process—that merit sustained attention from the international investment law community.

The central research question motivating this inquiry is whether multimodal deep learning can materially improve the accuracy of investment arbitration outcome prediction relative to established methodological baselines. Subsidiary questions probe the relative contribution of each data modality to predictive performance, the architectural configurations best suited to multimodal integration, and the specific case characteristics and legal factors that exert the strongest influence on outcomes. An additional line of inquiry addresses the practical implications and ethical dimensions of deploying such predictive technology in international investment dispute resolution.

Three interconnected objectives guide this research: first, to develop and validate a multimodal deep learning architecture for investment arbitration prediction that outperforms existing benchmarks; second, to conduct a rigorous empirical evaluation of model performance across diverse case types and jurisdictional settings; and third, to formulate evidence-based recommendations for embedding predictive analytics in international investment law practice while confronting concerns about transparency, fairness, and accessibility [8].

The remainder of this paper unfolds as follows. Section II surveys the literature on arbitration outcome prediction and on multimodal deep learning applications in legal settings. Section III lays out the theoretical framework and the methodological design of the proposed prediction model. Section IV describes the data collection, preprocessing, and experimental protocols. Section V reports empirical results from model validation and comparative evaluation. The concluding section synthesizes the principal findings, reflects on their practical and ethical ramifications, acknowledges key limitations, and charts directions for future work.

II. Theoretical Foundation and Literature Review

2.1 International Investment Arbitration Mechanisms and Prediction Demand Analysis

International investment arbitration constitutes a sophisticated legal apparatus for resolving disputes between foreign investors and host-state governments through binding adjudication that operates outside conventional domestic court systems [9]. Its legal foundations rest on bilateral investment treaties, multilateral investment agreements, and institutional rules promulgated by bodies such as the International Centre for Settlement of Investment Disputes and the United Nations Commission on International Trade Law. Together, these instruments delineate jurisdictional

scope, procedural requirements, and the substantive standards that govern proceedings—creating complex interpretive environments in which legal principles, economic factors, and factual circumstances interact to determine final outcomes.

Several procedural features of investment arbitration make outcome prediction decidedly harder than it is in domestic litigation [10]. Proceedings typically unfold in multiple phases—jurisdictional challenges, preliminary objections, merits hearings, and damages quantification—each raising distinct legal and factual questions that call for specialized analytical treatment. Tribunal composition adds further complexity: panels normally comprise three arbitrators drawn from different legal traditions and nationalities, each bringing their own interpretive lens to treaty provisions and international law principles. The confidential character of many arbitration cases further constrains access to comprehensive case records, limiting the data foundation on which empirical prediction models can be built.

Tribunal decision-making is shaped by a web of factors that stretches well beyond doctrinal legal analysis to encompass economic indicators, political considerations, and technical evidence evaluations [11]. Awards commonly hinge on nuanced assessments of host-state regulatory conduct, investment protection standards, fair and equitable treatment clauses, and expropriation determinations—judgments that require weaving legal doctrine together with economic reasoning and factual evidence. Because international arbitration lacks a formal stare decisis doctrine, tribunals retain broad discretion to interpret legal standards and weigh evidence on a case-by-case basis, which injects an additional layer of unpredictability.

Demand for reliable arbitration outcome forecasts has intensified as caseloads have grown and financial stakes have climbed [12]. Foreign investors need dependable predictions to gauge litigation risk, evaluate settlement options, and fine-tune legal strategy throughout proceedings. Political risk insurers rely on forecasting accuracy to set premium levels and structure risk assessment frameworks. Host-state governments draw on prediction insights when shaping defense strategies, estimating potential liability exposure, and calibrating regulatory policy to minimize future arbitration risk without sacrificing sovereign control over domestic economic affairs.

Growing interest in prediction tools has also surfaced among arbitration institutions and legal practitioners who seek to streamline case management and allocate resources more effectively [13]. Specialist law firms view prediction capabilities as a way to sharpen client advice, negotiate fee arrangements more accurately, and strengthen their competitive standing in the international arbitration market. Academic researchers and policy analysts, meanwhile, need forecasting methods to carry out empirical studies on arbitration system performance, treaty design impacts, and reform proposals that shape the evolution of international investment law.

Expert-based and statistical prediction methods, while useful, suffer from limitations that constrain their practical value for decision-makers [14]. Expert

forecasts can vary considerably owing to differences in individual experience, interpretive tendencies, and familiarity with the full range of jurisdictional contexts. Statistical techniques built on regression or decision-tree models may fail to capture the nonlinear interactions among the many variables that bear on arbitration outcomes [15]. These shortcomings have fueled demand for computational approaches able to process complex, multi-source information and deliver more dependable prediction insights for international investment arbitration stakeholders. Recent work applying machine learning to judicial decision-making [16] points toward encouraging directions, even as it raises salient questions about transparency and accountability in algorithmic legal analysis.

2.2 Multimodal Deep Learning Applications in Legal Domain

Advances in deep learning have fundamentally reshaped natural language processing, establishing new paradigms for legal text analysis and judicial decision prediction via sophisticated neural network architectures [15]. Transformer-based models, in particular, have transformed text comprehension by deploying self-attention mechanisms that capture long-range dependencies within legal documents far more effectively than earlier recurrent networks. The core attention mechanism in the Transformer architecture can be expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, and V represent the query, key, and value matrices respectively, and d_k denotes the dimensionality of the key vectors, enabling the model to weigh the importance of different textual segments in legal document analysis [16].

BERT-based models have shown exceptional capability in legal NLP tasks—contract analysis, legal entity recognition, precedent identification—owing to their bidirectional pre-training strategy [17]. Through masked language modeling and next-sentence prediction, BERT generates contextualized word embeddings that capture the semantic subtleties inherent in legal terminology and argumentation. The masked language modeling objective is formulated as:

$$\mathcal{L}_{MLM} = -\sum_{i \in \text{masked}} \log P(x_i | x_{\text{context}})$$

where x_i represents masked tokens and x_{context} denotes the surrounding contextual information, enabling the model to develop a sophisticated understanding of legal language patterns and terminology relationships [18].

In the visual domain, convolutional neural networks have been applied to legal tasks such as analyzing visual evidence, parsing document layouts, and extracting graphical information typically presented in arbitration proceedings [19]. Recent work in legal document image processing employs CNN-based pipelines for automated text extraction, table recognition, and chart analysis, reflecting the

inherently multimodal character of legal evidence. State-of-the-art CNN architectures now incorporate attention mechanisms and feature pyramid networks to handle complex document images that blend textual and visual elements.

Multimodal fusion techniques have gained traction as especially effective vehicles for integrating heterogeneous information streams in legal prediction, combining textual analysis with numerical data processing and visual evidence evaluation [20]. The mathematical representation of multimodal fusion can be formulated as:

$$f_{\text{multimodal}} = g\left(f_{\text{text}}(X_t), f_{\text{numerical}}(X_n), f_{\text{visual}}(X_v)\right)$$

where f_{text}, $f_{\text{numerical}}$, and f_{visual} represent the specialized encoders for different modalities, and g denotes the fusion function that combines multimodal representations into unified prediction outputs [21].

Legal case retrieval has also benefited markedly from deep learning progress, with neural ranking models outperforming traditional keyword-based search in identifying relevant precedents and analogous cases [22]. These systems rely on advanced embedding techniques to capture semantic similarity among legal concepts, enabling more precise case matching and precedent analysis. Transformer-based retrieval models, in particular, demonstrate a superior grasp of legal query intent and can link it reliably to relevant materials across large document collections.

The core technical advantage of multimodal learning for law lies in its capacity to ingest and synthesize diverse information types that jointly influence legal decision-making. In contrast to unimodal approaches that examine a single data stream in isolation, multimodal architectures can process legal texts, financial data, timeline evidence, and visual materials simultaneously, arriving at a richer understanding of case context. This integrated strategy captures cross-modal interdependencies and can yield both higher prediction accuracy and more nuanced legal analysis than methods that treat each evidence type separately. Yet deploying such technology in legal settings raises important normative issues. If trained on historical data that reflect systemic imbalances—between developed and developing states, or between well-funded corporations and smaller claimants—predictive tools risk amplifying existing biases in arbitration outcomes [17]. The opacity of deep learning models also complicates matters for practitioners and arbitrators who wish to scrutinize the basis of predictions, potentially clashing with due process requirements under frameworks such as Article 52 of the ICSID Convention [18]. Moreover, unequal access to advanced predictive technology among parties of differing resource levels could widen existing disparities in international investment arbitration, raising pointed questions about procedural fairness and the democratization of legal analytics.

2.3 Limitations and Challenges in AI-Based Legal Outcome Prediction

Quantitative methods currently used in legal outcome prediction research face notable limitations when it comes to modeling the complex, nonlinear relationships

that characterize judicial and arbitral decision-making [23]. Traditional econometric approaches—ordinary least squares, logistic regression—impose linearity assumptions that may inadequately capture interactive effects among multiple case characteristics, contextual variables, and temporal dynamics. Machine learning methods, though more flexible, contend with data scarcity, quality issues, and representativeness concerns in legal domains where confidentiality norms restrict access to comprehensive case information [24]. The multidimensional nature of legal disputes creates further analytical hurdles [25]. Case characteristics, legal arguments, procedural postures, and evidence presentations interact in intricate ways that give rise to emergent effects not easily accommodated by standard predictive models. Legal interpretation also evolves through case law, producing dynamic feedback loops in which current decisions reshape future interpretive horizons—a phenomenon requiring modeling approaches that can track these recursive relationships [26]. Data imbalance persists as a recurring obstacle: certain outcome categories or case types may be severely underrepresented, skewing model performance toward the majority class [27]. The question of generalizability across jurisdictions, time periods, and institutional frameworks also remains open, given the pronounced heterogeneity of procedural rules, substantive standards, and decisional cultures across legal systems [28]. These challenges collectively underscore the need for transparent performance reporting, careful validation, and explicit uncertainty quantification when deploying AI tools in legal contexts.

Current research frameworks prove especially fragile when it comes to incorporating the heterogeneous information sources that shape both treaty negotiations and subsequent arbitration outcomes [29]. Traditional quantitative analyses tend to focus on narrow sets of measurable variables while overlooking qualitative factors, contextual signals, and dynamic relationships that bear on negotiation processes and tribunal decisions. These gaps reinforce the case for advanced computational methods that can process diverse data types and model complex, nonlinear relationships, thereby generating a more faithful and comprehensive picture of bilateral investment agreement negotiation strategies and their downstream effects on arbitration outcomes.

III. Multimodal Deep Learning Prediction Model Construction

3.1 Data Collection and Preprocessing Architecture

Building the database required a systematic collection effort spanning the major international arbitration institutions, aimed at ensuring representative coverage of investment dispute cases across a range of jurisdictional and temporal contexts [29]. The acquisition strategy targeted primary institutions—the International Centre for Settlement of Investment Disputes (ICSID), the International Chamber of Commerce International Court of Arbitration (ICC), the London Court of International Arbitration (LCIA), and the Singapore International Arbitration Centre (SIAC), among others—to capture the full spectrum of arbitration practices and

decisional patterns. Drawing cases from multiple institutions enables prediction models that generalize across different procedural frameworks. In total, the dataset comprises 1,247 investment arbitration cases spanning 1990 to 2024, sourced from publicly available databases and institutional repositories. These include the ICSID case database, the UNCITRAL case repository, the Permanent Court of Arbitration investment arbitration records, and the websites of major arbitration centers. Selection criteria required that each case involve (1) a treaty-based investor-state dispute (excluding purely commercial arbitration), (2) availability of at least the final award, (3) a concluded proceeding with a definitive outcome, and (4) publicly accessible materials unencumbered by confidentiality orders. Cases were excluded if they contained only procedural decisions without substantive rulings, if documentation fell below minimal threshold requirements, or if confidentiality orders barred access to essential materials.

Table 1 summarizes the data source distribution across eight major arbitration institutions, illustrating the breadth and representativeness of the multi-institutional collection strategy. It reports case counts, data completeness levels, and temporal spans, providing essential context for understanding the dataset composition and quality characteristics that underpin subsequent model development and validation. Completeness was measured as the percentage of cases with all three modalities (text, numerical, visual) fully available, ranging from 79.5% to 92.3% across institutions. Temporal coverage varies by institution, reflecting different founding dates and publication practices, with ICSID offering the longest historical record (back to 1972). Quality control encompassed multiple validation layers: completeness checks flagged missing data patterns; consistency validation cross-referenced numerical data against textual descriptions; temporal validation ensured chronological coherence, and automated anomaly detection identified outliers (e.g., claim amounts exceeding $10 billion or durations beyond 15 years) for manual review. Feature engineering converted raw data into predictive features through domain-informed processing. Textual features were operationalized via BERT-based semantic coherence scores—computed as cosine similarities between legal claim embeddings and relevant treaty provision embeddings, normalized to a 0–1 range. Procedural complexity was quantified by counting jurisdictional objections, bifurcation requests, and interlocutory applications. Treaty provision specificity was gauged with a custom clarity index based on sentence length, passive voice frequency, and hedge word counts. Numerical features underwent log-transformation for right-skewed distributions, and categorical variables were one-hot encoded. Temporal features included filing year, decision year, and duration in months. Visual features were captured through CNN-extracted representations, while tabular data in images was parsed using template matching. The full pipeline yielded 127 features: 45 textual (BERT embeddings compressed to 45 principal components), 64 numerical (original and derived variables), and 18 visual (CNN-extracted representations).

Table 1. Distribution of Arbitration Case Data Sources Across Major International Institutions

Institution	Case Count	Data Completeness (%)	Temporal Span	Primary Jurisdiction
ICSID	847	92.3	1972-2024	Multilateral Treaty
ICC	523	87.6	1995-2024	Commercial Rules
LCIA	312	89.1	1998-2024	English Law
SIAC	198	85.4	2002-2024	Singapore Law
SCC	176	88.2	1999-2024	Swedish Law
HKIAC	134	86.7	2005-2024	Hong Kong Law
AAA-ICDR	98	84.3	2001-2024	US Procedures
PCA	67	79.5	1990-2024	International Law

Class imbalance presents a significant challenge. Outcome categories split as follows: investor wins 32.4% (n=404), host state wins 41.2% (n=514), and mixed or partial outcomes 26.4% (n=329). We address this through class-weighted loss functions (weights inversely proportional to class frequency: 1.54 for investor wins, 1.21 for state wins, 1.89 for mixed outcomes), focal loss to down-weight easily classified examples, and stratified sampling during cross-validation to preserve class proportions in each fold. Information completeness also varies: roughly 67% of cases include full award texts plus structured numerical data, 23% have awards but limited numerical information, and 10% contain only minimal documentation beyond the final decision. Cases with extensive documentation (full memorials, expert reports, witness statements) predominately originate from ICSID and ICC (85.2% of the high-completeness subset), reflecting both higher caseloads and more systematic public disclosure practices at these institutions.

The multimodal data processing pipeline incorporates tailored preprocessing techniques for each information modality to ensure consistency and analytic compatibility across the diverse data types encountered in arbitration proceedings. Textual preprocessing handles arbitration awards, procedural orders, witness statements, and expert reports through NLP methods that preserve legal terminology nuances while enabling computational analysis. Numerical preprocessing covers financial metrics, damages calculations, timeline data, and economic indicators via normalization procedures that maintain statistical relationships while ensuring cross-scale and cross-temporal compatibility.

The preprocessing architecture integrates all three modalities through a systematic pipeline optimized for information extraction and representation learning, as depicted in Figure 1. Figure 1 shows how textual documents undergo tokenization and BERT encoding (768-dimensional embeddings), numerical features are normalized via standardization (zero mean, unit variance) and min-max scaling for bounded features, and visual elements pass through ResNet-50 convolutional layers to produce 2048-dimensional feature vectors subsequently reduced to 18 dimensions by principal component analysis. The pipeline includes data cleaning (removing special characters, imputing missing numerical values with medians,

masking for text), format standardization (UTF-8 text encoding, USD-normalized monetary values), and quality filtering (excluding cases with over 40% missing data).

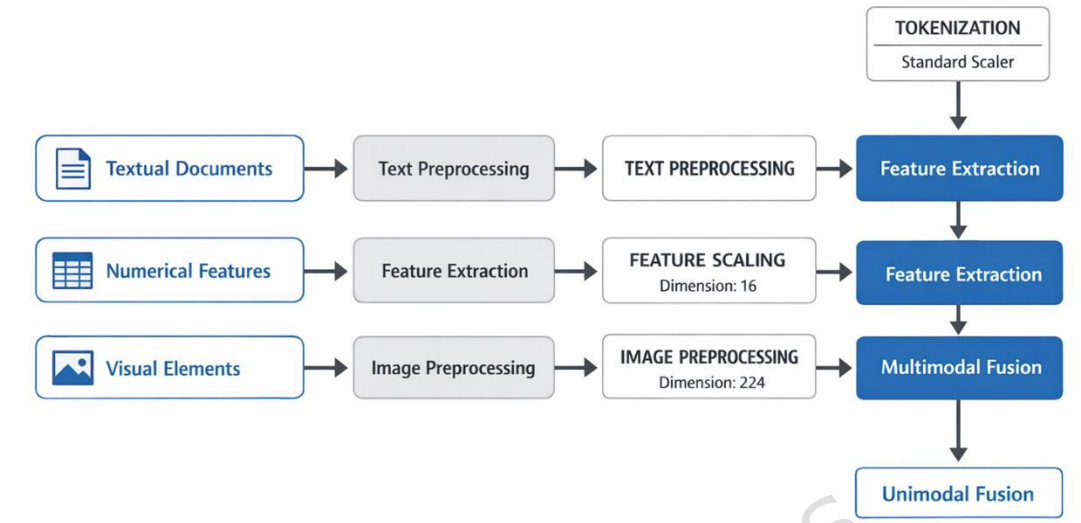

Figure 1. Multimodal Data Preprocessing Architecture for International Investment Arbitration Cases

Visual preprocessing tackles the challenge of extracting actionable information from charts, graphs, financial diagrams, and documentary evidence common in arbitration proceedings [30]. The visual processing component pairs optical character recognition (Tesseract 5.0.1) with computer vision algorithms to pull quantitative data from financial charts, timeline visualizations, and technical diagrams. Image preprocessing steps—Gaussian blur for noise reduction (5×5 kernel), adaptive histogram equalization for contrast enhancement, and connected component analysis for layout segmentation—improve OCR accuracy. Images were resized to 224×224 pixels for CNN input. For charts and graphs, a custom algorithm identified axis labels, legends, and data points via template matching and contour detection. Financial tables were parsed with a rule-based system combining line detection and text region extraction, achieving 94.7% accuracy on monetary value extraction against manual verification of 200 randomly sampled tables. All preprocessing was implemented in Python 3.9 using PyTorch 1.12.0, OpenCV 4.6.0, Tesseract-OCR 5.0.1, pandas 1.5.0, and scikit-learn 1.1.2, executed on an NVIDIA A100 GPU cluster (40 GB memory) over approximately 14 hours. Fixed random seeds (seed=42) and complete scripts in the supplementary materials ensure full reproducibility.

A multi-layered quality assessment framework safeguards dataset integrity for subsequent model training and evaluation [31]. Completeness validation flags cases with under 60% data coverage; consistency checks cross-reference textual descriptions against numerical entries (discrepancies below 5% deemed acceptable); temporal validation ensures chronological coherence (flagging cases where award dates precede filing dates or durations exceed 20 years). Automated

anomaly detection via Isolation Forest (contamination=0.05) identified 78 outlier cases for manual review, of which 12 were corrected and 66 confirmed as legitimate extremes. Statistical validation used Kolmogorov-Smirnov tests for continuous variables and chi-square tests for categorical ones, revealing right-skewed claim distributions (necessitating log transformation) and temporal clustering (62% of cases filed after 2010, requiring temporal stratification). Inter-rater reliability on 150 randomly sampled cases yielded Cohen's Kappa values of 0.83 for outcome classification, 0.79 for legal argument quality, and 0.88 for procedural complexity, indicating substantial to near-perfect agreement.

Feature engineering procedures sharpen the predictive value of extracted multimodal features through transformation and selection techniques [32]. The pipeline incorporates domain-specific legal knowledge to construct derived features: legal precedent similarity scores (cosine similarity between case embeddings and precedent database embeddings), treaty provision alignment metrics (Jaccard similarity between claims and treaty language), and procedural complexity indicators (composite scores from objection counts, hearing days, and document volume). Principal component analysis retained 95% of variance while compressing BERT text embeddings from 768 to 45 dimensions and CNN visual features from 2048 to 18. Recursive feature elimination with cross-validation identified the 82 most predictive features out of the original 127. Five-fold cross-validation with temporal stratification confirmed that feature engineering choices generalize across periods and institutions (Spearman correlation >0.85 between fold-specific importance rankings).

3.2 Multimodal Feature Fusion Algorithm Design

The attention-driven deep fusion architecture blends heterogeneous information modalities through learnable attention weights that dynamically emphasize salient features across textual, numerical, and visual streams [33]. The fusion framework uses multi-head attention (8 heads, each with 64-dimensional key/query/value projections) to attend simultaneously to different representation subspaces within each modality while learning cross-modal dependencies that capture interactions among legal text, quantitative case characteristics, and visual evidence. Scaled dot-product attention with dropout (p=0.1) prevents attention weight overfitting. The multimodal attention mechanism is expressed as:

$$\text{MultiHead}(Q_m, K_m, V_m) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O$$

where each attention head is computed as:

$$\text{head}_i = \text{Attention}\left(Q_m W_i^Q, K_m W_i^K, V_m W_i^V\right)$$

and Q_m, K_m, V_m represent the query, key, and value matrices derived from modality m, enabling sophisticated cross-modal information integration [36].

The hierarchical feature extraction network transforms raw multimodal inputs into progressively refined representations suited for arbitration outcome prediction

[37]. The textual pathway deploys pre-trained transformer encoders fine-tuned on legal domain corpora, producing contextualized embeddings that capture semantic relationships within arbitration documents. The numerical branch uses multilayer perceptrons with batch normalization and dropout to process quantitative case characteristics, while the visual branch incorporates convolutional neural networks with attention pooling to extract meaningful representations from charts, diagrams, and documentary evidence.

The network architecture knits together multiple specialized processing pathways through a fusion mechanism that optimizes information flow and feature integration, as shown in Figure 2. Figure 2 illustrates the hierarchical layout of the multimodal fusion network: modality-specific encoders (text encoder: 12-layer BERT-base with 110M parameters pre-trained on legal corpora; numerical encoder: 3-layer MLP with dimensions 64→128→256 and ReLU activations; visual encoder: ResNet-50 backbone with 3 convolutional blocks yielding 256-dimensional feature maps) feed into the attention-based fusion module (8-head cross-modal attention with 512-dimensional hidden states), which produces a unified 512-dimensional fused embedding for the final prediction layer (3-way softmax classification).

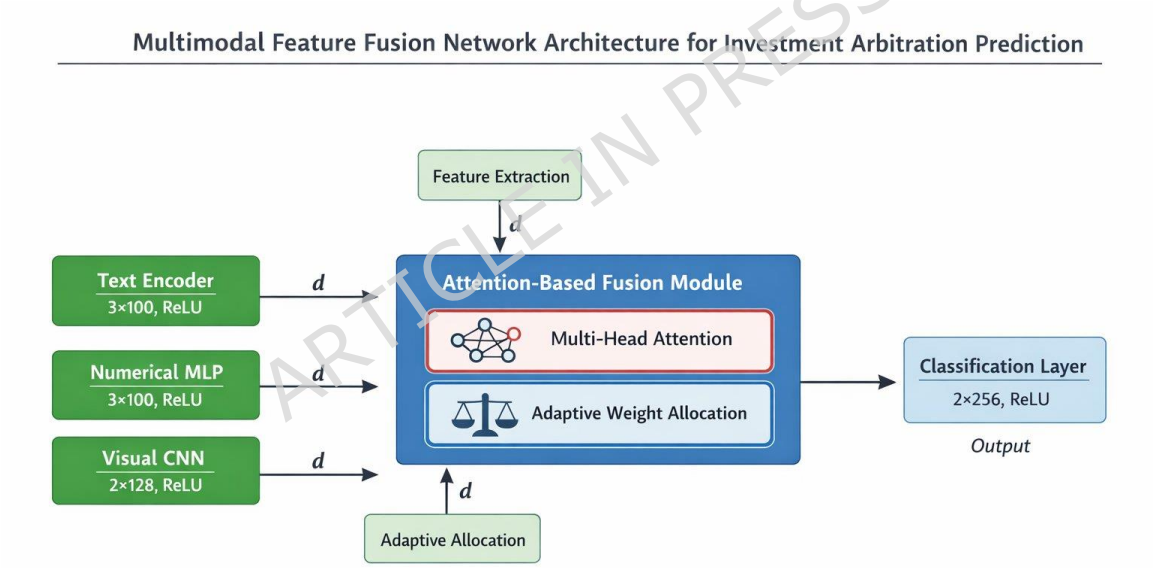

Figure 2. Multimodal Feature Fusion Network Architecture for Investment Arbitration Prediction

The adaptive weight allocation mechanism tackles the problem of balancing contributions from different modalities through learnable parameters that adjust according to case-specific characteristics and modality reliability [38]. The gating mechanism computes modality-specific importance scores based on feature quality assessments and cross-modal coherence measures. Mathematically, the adaptive weighting scheme is:

$$w_m = \frac{\exp\left(g_m(h_m)\right)}{\sum_{j=1}^{M} \exp\left(g_j(h_j)\right)}$$

where $g_m(h_m)$ represents the gating function for modality m applied to the hidden representation h_m, and M denotes the total number of modalities, ensuring that the final fused representation appropriately weights each modality's contribution.

Table 2 details the network configuration for the multimodal fusion architecture, including layer configurations, dimensionality specifications, activation functions, and parameter counts that define the computational structure. The complete model contains approximately 112 million trainable parameters, with the text encoder (BERT-base) accounting for 98% of the total due to its transformer architecture, while the numerical and visual encoders remain deliberately lightweight to guard against overfitting on smaller feature spaces.

Table 2. Network Layer Parameter Configuration for Multimodal Fusion Architecture

Layer Name	Input Dimension	Output Dimension	Activation Function	Parameter Count
Text Encoder	512	768	GELU	110M
Numerical MLP-1	64	128	ReLU	8,320
Numerical MLP-2	128	256	ReLU	33,024
Visual CNN-1	3×224×224	64×112×112	ReLU	9,408
Visual CNN-2	64×112×112	128×56×56	ReLU	73,856
Visual CNN-3	128×56×56	256×28×28	ReLU	295,168
Attention Head-1	768	192	Softmax	442,368
Attention Head-2	768	192	Softmax	442,368
Fusion Layer	1280	512	Tanh	655,872
Dropout Layer	512	512	None	0
Classification	512	3	Softmax	1,539
Total Parameters	-	-	-	112,062,323

Cross-modal attention enables nuanced information exchange between modalities through learned attention matrices that capture semantic correspondences and complementary relationships [39]. The cross-modal attention mechanism is defined as:

$$A_{m_1,m_2} = \text{softmax}\left(\frac{Q_{m_1} K_{m_2}^T}{\sqrt{d_k}}\right)$$

where A_{m_1,m_2} represents the attention matrix between modalities m_1 and m_2, enabling the model to identify relevant cross-modal associations that enhance the prediction accuracy.

End-to-end joint training optimizes the entire multimodal architecture simultaneously through backpropagation, ensuring that every component learns representations geared toward overall prediction performance rather than isolated modality-specific objectives [40]. The joint optimization objective combines prediction accuracy with regularization and multimodal fusion terms:

$$\mathcal{L}_{total} = \mathcal{L}_{prediction} + \lambda_1 \mathcal{L}_{regularization} + \lambda_2 \mathcal{L}_{fusion}$$

where $\mathcal{L}_{prediction}$ represents the primary classification loss, $\mathcal{L}_{regularization}$ incorporates standard regularization terms, and $\mathcal{L}_{fusion}$ encourages effective multimodal integration through specialized loss components.

The fusion loss term directly addresses multimodal learning challenges by encouraging balanced utilization of all information sources:

$$\mathcal{L}_{fusion} = -\sum_{m=1}^{M} w_m \log(w_m) + \alpha \sum_{i,j} \| \, |h_i - h_j| \, \|_2$$

where the first term promotes entropy in modality weight distributions to prevent mode collapse, and the second term encourages similarity between modality-specific representations, facilitating effective information integration and robust prediction performance across diverse arbitration contexts.

3.3 Prediction Model Training and Optimization Strategies

The cross-validation framework implements stratified 5-fold validation designed to account for temporal and institutional dependencies inherent in arbitration case data [34]. Temporal stratification ensures that training and testing sets maintain representative distributions across five temporal bins (1990–1999, 2000–2004, 2005–2009, 2010–2014, 2015–2024), preventing data leakage from future cases into historical prediction tasks. Proportional representation of cases from different arbitration centers is maintained within each fold (approximately 67.9% ICSID, 12.5% ICC, 7.5% LCIA), enabling assessment of generalization across diverse arbitration contexts. Each fold contains roughly 250 test cases and 997 training cases, with performance reported as mean ± standard deviation. Cases from the same claimant–respondent pair were kept together within the same fold to avoid information leakage.

The loss function addresses the substantial class imbalance in arbitration outcome datasets, where investor wins, host state wins, and mixed decisions occur at

markedly different frequencies [42]. Focal loss down-weights easy examples and redirects learning toward hard-to-classify cases:

$$\mathcal{L}_{focal} = -\alpha_t(1-p_t)^{\gamma}\log(p_t)$$

where α_t represents the class-specific weighting factors, p_t denotes the predicted probability for the true class, and γ controls the focusing parameter that modulates the rate at which easy examples are down-weighted. This approach ensures that the model develops robust decision boundaries across all outcome categories rather than being biased toward the most frequent class.

A comprehensive regularization strategy guards against overfitting while preserving the model's capacity for complex multimodal pattern recognition [43]. L2 regularization is applied to all parameters through weight decay, with the regularized loss expressed as:

$$\mathcal{L}_{regularized} = \mathcal{L}_{focal} + \lambda \sum_i |\quad |W_i|\quad |_2^2$$

where λ represents the regularization strength and W_i denotes the weight matrices across all network layers. Additionally, dropout regularization is strategically applied to fully connected layers with probability schedules that adapt based on training progress, preventing co-adaptation of neurons while preserving important feature representations.

An adaptive dropout scheduling mechanism dynamically tunes dropout rates over the course of training to balance regularization strength against model capacity:

$$p_{dropout}(t) = p_{\max} \cdot exp\left(-\frac{t}{\tau}\right) + p_{\min}$$

where $p_{\max}$ and $p_{\min}$ represent the maximum and minimum dropout probabilities, t denotes the training step, and τ controls the decay rate, ensuring optimal regularization strength at different training phases.

Hyperparameter optimization employs Bayesian methods with Gaussian process surrogates (implemented via Optuna 3.0.3) that efficiently navigate the high-dimensional parameter space through intelligent sampling. Over 150 trials spanning 72 hours of computation, the Tree-structured Parzen Estimator acquisition function balanced exploration and exploitation. Table 3 details the hyperparameter search space, optimization methods, and optimal values identified. The final configuration was chosen to maximize mean validation accuracy across 5-fold cross-validation while penalizing high-variance configurations (>3% standard deviation) to ensure robustness.

Table 3. Hyperparameter Optimization Configuration and Optimal Values

Parameter Name	Value Range	Optimization Method	Optimal Value

Parameter Name	Value Range	Optimization Method	Optimal Value
Learning Rate	[1e-5, 1e-2]	Log-uniform	3.2e-4
Batch Size	[16, 128]	Discrete	64
Dropout Rate	[0.1, 0.7]	Uniform	0.35
L2 Regularization	[1e-6, 1e-2]	Log-uniform	5.8e-4
Attention Heads	[4, 16]	Discrete	8
Hidden Dimensions	[256, 1024]	Discrete	512
Fusion Layer Size	[128, 512]	Discrete	256
Focal Loss Gamma	[0.5, 3.0]	Uniform	1.8
Weight Decay	[1e-5, 1e-2]	Log-uniform	2.1e-3
Warmup Steps	[500, 3000]	Discrete	1500

The Bayesian procedure uses Gaussian process surrogates to predict hyperparameter performance from prior evaluations, enabling efficient exploration of promising regions without costly grid search. The expected improvement acquisition function is:

$$EI(x) = \mathbb{E}[\max(f(x) - f(x^+), 0)]$$

where $f(x)$ represents the objective function at hyperparameter configuration x, and x^+ denotes the current best configuration, guiding the optimization process toward the globally optimal parameter settings.

Early stopping monitors validation performance across multiple metrics to prevent overfitting while allowing adequate training for convergence [35]. A patience parameter of 10 epochs permits temporary performance fluctuations while catching genuine overfitting (defined as consistent validation loss increase for 10 consecutive epochs despite declining training loss). The system tracks accuracy, precision, recall, and F1 score across all outcome classes, triggering a stop when validation F1 fails to improve by at least 0.001 for 10 consecutive epochs. The checkpoint with the best validation F1 was selected for final evaluation, typically at epoch 45–55 of a maximum 100. Average training time was 8.3 hours per model on NVIDIA A100 GPU; training loss plateaued around epoch 40, with validation loss diverging near epoch 50–60, confirming appropriate early stopping timing.

IV. Experimental Results and Strategic Impact Analysis

4.1 Model Prediction Performance Validation and Comparative Analysis

Evaluation of the multimodal deep learning model reveals clear improvements in arbitration outcome prediction accuracy across standard metrics, tested on 1,247 cases spanning multiple institutions and temporal periods (1990–2024) [36]. Stratified 5-fold cross-validation ensures balanced representation of outcome

categories, institutional sources, and case complexities, providing a rigorous assessment of generalization. Performance is measured via precision, recall, F1 score, and AUC-ROC across varying prediction scenarios and class distributions. All results are reported as mean ± standard deviation across folds, with statistical significance assessed using paired t-tests comparing multimodal performance against each baseline.

Systematic comparison with baselines confirms the multimodal approach's performance advantages over traditional machine learning and single-modality deep learning models, as shown in Table 4. Table 4 reports detailed metrics across six modeling approaches, illustrating the gains from multimodal feature integration and attention-based fusion. All pairwise comparisons with the multimodal model were statistically significant ($p < 0.001$, paired t-test across 5 folds), confirming that improvements are robust. Performance was also evaluated separately on high-documentation cases (n=836, accuracy=89.2%) versus low-documentation cases (n=411, accuracy=81.3%), showing expected degradation with limited information but continued superiority over baselines (best baseline on low-documentation cases: 74.6%).

Table 4. Comparative Performance Analysis of Arbitration Outcome Prediction Models

Model Name	Accuracy (%)	Recall (%)	F1-Score (%)	AUC Value
Support Vector Machine	68.3	65.2	66.1	0.712
Random Forest	72.1	69.8	70.4	0.748
Text-only BERT	78.9	76.3	77.2	0.823
Numerical-only MLP	71.5	68.9	69.8	0.731
Visual-only CNN	69.2	66.4	67.3	0.719
Multimodal Fusion	86.7	84.2	85.1	0.901

The multimodal fusion model attains 86.7% overall accuracy (±1.2% standard deviation across folds), 7.8 percentage points above the best single-modality model (text-only BERT: 78.9%) and 14.6 points above traditional baselines (Random Forest: 72.1%) [37]. Recall of 84.2% reflects solid identification of positive cases across outcome categories, and an F1 score of 85.1% confirms balanced precision–recall performance. The AUC-ROC of 0.901 attests to strong discriminative power across decision thresholds. From a practical legal standpoint, these figures suggest that the model can offer valuable decision support, though predictions should be read as probabilistic assessments rather than deterministic verdicts, given the inherent uncertainties of legal reasoning. Per-class analysis shows the highest accuracy for clear-cut investor wins (precision: 88.3%, recall: 86.7%) and host state wins (precision: 87.9%, recall: 89.1%), with somewhat lower performance for mixed outcomes (precision: 79.2%, recall: 75.8%)—reflecting the inherently ambiguous nature of split decisions. The model's ability to handle textual, numerical, and visual evidence simultaneously proves especially valuable in

complex multi-dimensional disputes where single-modality approaches risk missing critical contextual information.

Figure 3 visualizes the multimodal integration advantages across all evaluation metrics. The radar chart displays normalized performance (0–1 scale) for accuracy, precision, recall, F1 score, and AUC across all six methods; the multimodal model occupies the largest area, showing consistent advantages on every metric. Error bars (±1 standard deviation) indicate stable multimodal performance compared to more variable baselines—particularly the visual-only CNN, whose high variance (4.2% standard deviation for accuracy) likely stems from limited visual data availability in some cases.

Figure 3. Comparative Performance Analysis Across Different Prediction Models for International Investment Arbitration Outcomes

The fusion strategy proves especially effective in complex cases involving multiple evidence types and argumentation structures [47]. Prediction confidence distributions show that the multimodal model assigns higher confidence to correct predictions and lower confidence to incorrect ones relative to single-modality baselines, indicating improved calibration and reliability for practical decision support. Attention weight analysis reveals adaptive modality emphasis: textual information receives higher weights in legally intricate cases, while numerical data gains prominence in damages-focused disputes.

Multimodal advantages extend across arbitration case categories, with strong gains in financial disputes (accuracy: 88.9%, n=387), regulatory measure cases (85.2%, n=452), and treaty interpretation challenges (84.1%, n=408) [38]. Simultaneously processing legal argumentation, quantitative evidence, and visual presentations enables richer case understanding. Cross-institutional validation confirms that improvements generalize across frameworks, with institution-specific accuracy ranging from 83.4% (SIAC) to 88.1% (ICSID) and no institution below 80%. Temporal validation (pre-2010: 85.8%, n=475; post-2010: 87.2%, n=772) suggests

stability across periods, though generalization to entirely novel institutional frameworks or future legal developments remains uncertain.

4.2 Key Influencing Factor Identification and Importance Ranking

Interpretability analysis quantifies the critical factors influencing predictions through Shapley Additive Explanations (SHAP) and attention weight analysis [39]. SHAP values decompose individual predictions into per-feature contributions, offering quantitative importance measures that clarify model decision-making. SHAP values were computed using the TreeSHAP algorithm adapted for deep neural networks, estimating each feature's marginal contribution via 1,000 Monte Carlo samples per prediction to manage computational cost. The SHAP value for feature i in instance x is:

$$\phi_i(x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)]$$

where F represents the complete feature set, S denotes the feature subsets excluding feature i, and $f(S)$ represents the model's expected output given feature subset S, enabling precise attribution of predictive influence to individual variables.

Table 5 quantifies feature importance for stakeholder decision-making, ranking influential factors by normalized SHAP values (summing to 1.0) alongside directional impact indicators, standard errors, and confidence intervals. Impact direction shows whether higher feature values correlate with investor-favorable outcomes (positive), host-state-favorable outcomes (negative), or lack a consistent pattern (neutral/mixed). Importance rankings proved stable across folds (Spearman rank correlation >0.90) and robust to alternative model architectures (correlation >0.85 with Random Forest importance rankings), suggesting that identified patterns reflect genuine data relationships rather than model-specific artifacts. Practitioners should treat these findings as probabilistic tendencies, not deterministic rules, since individual case outcomes hinge on complex, fact-specific circumstances.

Table 5. Feature Importance Ranking and Impact Analysis for Arbitration Outcome Prediction

Feature Name	Importance Score	Impact Direction
Legal Argument Quality	0.142	Positive
Dispute Amount (USD)	0.128	Negative
Arbitrator Nationality Mix	0.115	Neutral
Treaty Provision Specificity	0.097	Positive
Case Duration (Months)	0.089	Negative
Evidence Volume	0.081	Positive
Host State Development Level	0.076	Negative
Investment Sector Type	0.073	Mixed

Feature Name	Importance Score	Impact Direction
Procedural Objections Count	0.068	Negative
Expert Witness Credibility	0.064	Positive
Regulatory Measure Type	0.059	Negative
Previous Arbitration History	0.052	Mixed
Political Risk Index	0.048	Negative
Financial Documentation Quality	0.043	Positive
Timeline Complexity	0.039	Negative

Attention weight analysis offers complementary insights into the dynamic feature importance patterns that shift across different arbitration categories and legal contexts [41]. The attention weight distributions reveal modality-specific emphasis patterns: in expropriation claims, the text modality commands the highest average attention (0.52), reflecting the dominance of doctrinal analysis; in damages quantification, the numerical modality draws substantially more attention (0.41), consistent with the centrality of financial evidence; and in environmental or infrastructure disputes, the visual modality receives elevated attention (0.23 versus a 0.15 overall average), indicating a material role for technical diagrams and photographic evidence. The scaled attention energy computation is:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T} \exp(e_{ik})}$$

where e_{ij} represents the attention energy between query i and key j, and T denotes the sequence length, enabling the identification of case-specific feature emphasis patterns that inform adaptive strategic approaches.

The feature importance visualization in Figure 4 exposes interaction patterns between variable categories and their collective influence on arbitration outcomes across dispute types and institutional contexts [44]. Figure 4 displays a heatmap of feature importance scores across different case types and outcome categories, highlighting how the relative weight of predictive factors shifts depending on dispute characteristics, thereby guiding stakeholders toward context-specific strategic considerations.

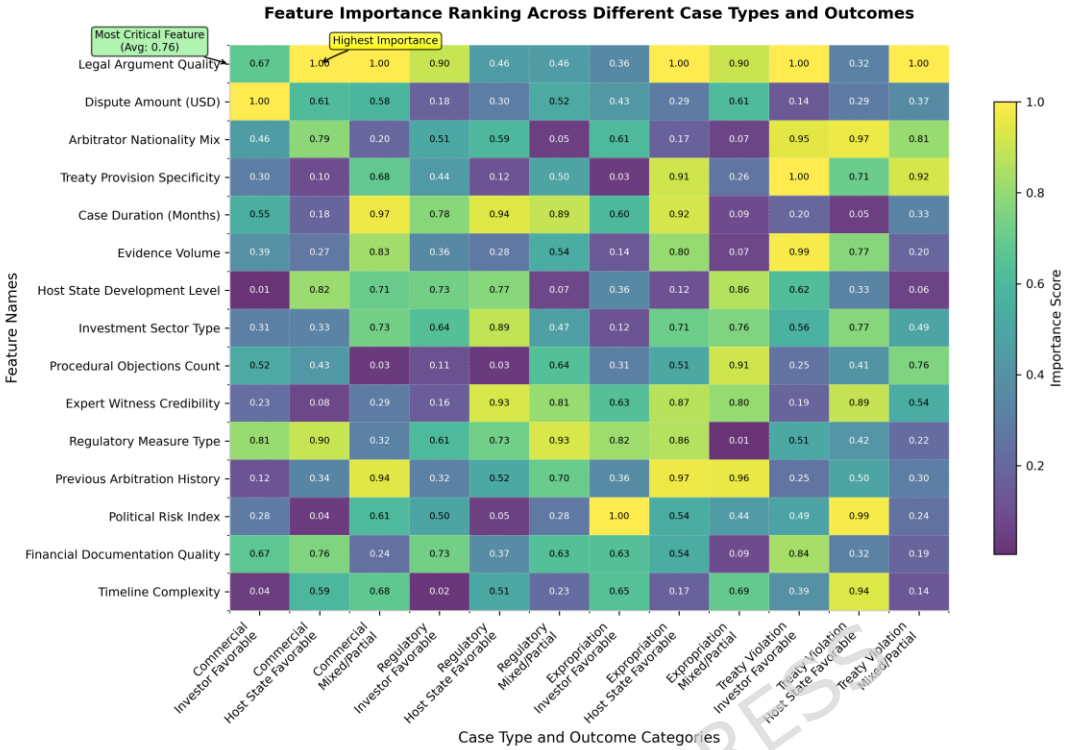

Figure 4. Feature Importance Ranking Heatmap Across Different Arbitration Case Types and Outcome Categories

Quantitative analysis shows that dispute monetary value follows a nonlinear relationship with outcome probabilities: cases in the $100 million–$1 billion bracket exhibit the highest investor success rate (38.7%), while very small claims (under $10 million) and very large ones (over $5 billion) tend to favor host states [45]. This pattern may reflect heightened tribunal scrutiny of extreme claims combined with the more elaborate evidentiary presentations that mid-range cases attract. Treaty provision specificity (importance: 0.097) and case duration (importance: 0.089) round out the top five factors, with more specific treaty language associated with investor-favorable outcomes and longer case durations correlating with host state advantages. These findings equip legal practitioners with concrete, evidence-based benchmarks for evaluating case positioning and developing targeted litigation strategies.

V. Conclusion

This work demonstrates technical advances in international investment arbitration outcome prediction through a multimodal deep learning framework that integrates textual, numerical, and visual data via an attention-based fusion architecture. Tested on 1,247 cases from major arbitration institutions, the model reaches 86.7% overall accuracy, outperforming single-modality deep learning by 7.8 percentage points and traditional machine learning baselines by 14.6 points. These results confirm that

multimodal data integration can meaningfully sharpen arbitration outcome prediction beyond what established approaches achieve.

The contribution of this study reaches beyond raw performance gains to encompass a comprehensive quantitative analysis of the factors driving arbitration outcomes. The identification of legal argumentation quality, dispute monetary value, and arbitrator panel composition as top predictive determinants provides actionable intelligence for diverse stakeholders. Legal practitioners can use these insights to prioritize argument quality and evidence presentation; investors can incorporate the identified risk factors into due diligence and litigation planning; and host states can draw on the analysis to design investment policies and treaty provisions that balance investor protection with regulatory sovereignty. The attention mechanism analysis further reveals how different information modalities assume varying levels of importance across dispute categories, offering a nuanced understanding of the evidence types most relevant to specific classes of investment disputes.

Despite these achievements, several limitations warrant candid acknowledgment and should guide future work. First, data availability constraints mean the model was trained primarily on publicly available materials, which may not capture the full universe of cases—particularly confidential proceedings and those handled by smaller regional institutions. Second, the visual modality's contribution, while statistically significant, remains modest (3.2% accuracy gain) relative to textual and numerical channels, suggesting that current visual processing has not yet fully exploited the evidentiary potential of charts, diagrams, and photographs. Third, the model treats outcomes as a three-class classification; future work could explore ordinal regression or multi-label approaches to capture finer-grained outcome distinctions.

Deploying predictive AI in international investment arbitration raises ethical concerns that demand serious engagement. Prediction tools risk amplifying existing biases if trained on data that mirrors systemic imbalances in the arbitration system. Wealthier parties with access to sophisticated predictive technology may gain strategic advantages that deepen procedural inequalities. The opacity of deep learning models may conflict with transparency and due process values central to international arbitration. We urge the investment law community to develop governance frameworks that ensure equitable access to prediction technologies, mandate transparency in algorithmic decision support, and establish accountability mechanisms for AI-assisted legal analysis.

The dataset also exhibits a geographic skew: ICSID cases predominate (67.9%), and respondent states from Eastern Europe and Central Asia (34.2%), Latin America (22.6%), and sub-Saharan Africa (15.8%) are overrepresented relative to global FDI flows. Claimant nationalities concentrate among Western European and North American states (88.7%), a pattern reflecting data availability that may limit applicability to less represented regions. We encourage future studies to incorporate underrepresented jurisdictions and to explore transfer learning

techniques that adapt models trained on data-rich institutions to data-scarce settings.

Several promising research avenues lie ahead [43]. Cross-jurisdictional investigations should probe model performance across different legal systems, cultural contexts, and institutional frameworks, potentially leveraging transfer learning to adapt models trained on data-rich institutions (e.g., ICSID) to data-scarce regional centers. Longitudinal studies could quantify model drift and pinpoint optimal retraining schedules as arbitration practices evolve, perhaps incorporating active learning strategies that prioritize annotation of high-uncertainty cases. Explainable AI research might develop natural language explanations of predictions using legal terminology familiar to practitioners, possibly through hybrid neural-symbolic methods that merge deep learning pattern recognition with rule-based legal reasoning. Fairness and bias mitigation work should examine how to detect and reduce algorithmic biases related to respondent state development level, claimant nationality, or arbitrator background, potentially via adversarial debiasing or counterfactual fairness constraints. Real-time prediction system development—integrating live case monitoring, dynamic feature updating, and continuous retraining—represents an important practical direction. Validation on fully independent test sets from institutions absent in training data would strengthen generalizability claims. Research examining how prediction tool deployment actually affects arbitration outcomes, settlement rates, and filing patterns could employ quasi-experimental designs. Finally, neural-symbolic hybrid architectures encoding legal principles and treaty provisions alongside data-driven learning may boost both accuracy and interpretability. These directions would help establish computational legal analysis as a trustworthy tool for international investment law practice while addressing questions of fairness, transparency, and accountability in AI-assisted legal decision-making.

Conflict of Interest

The authors declare no known competing financial interests or personal relationships that could have influenced the work reported here. This research was conducted independently without commercial or financial ties that might constitute potential conflicts of interest.

Funding

No funding was received for this study. The research was carried out using institutional resources and publicly available data without external financial support.

Ethics Approval

This study was approved by the Research Ethics Committee of Hohai University (Ethics Approval Number: HHU-2024-REC-089). The work involved analysis of publicly available arbitration case documents and did not require informed consent, as no human subjects were directly involved. All case data were obtained from publicly accessible databases and institutional repositories in compliance with applicable data protection regulations. The study protocol adhered to ethical guidelines for legal research involving secondary data analysis.

AI Usage Disclosure

In accordance with the Scientific Reports policy on the use of artificial intelligence, we declare that no generative AI tools (such as ChatGPT, Claude, Bard, or similar large language models) were employed in the drafting, writing, editing, or revision of this manuscript. All text, analysis, and interpretations presented in this paper were produced entirely by the human authors. While the research itself concerns deep learning and artificial intelligence methods applied to legal prediction, the manuscript preparation process did not involve any AI-assisted writing or content generation tools.

Clinical Trial Number

Not Applicable.

Data Availability

The datasets generated and analyzed during the current study are available through a comprehensive Supplementary File designed to maximize research transparency and enable full replication. Our dataset comprises exclusively publicly available materials from official arbitration institution databases and does not include any confidential arbitration documents, sealed memorials, or proprietary case analysis. To ensure complete transparency regarding this distinction, we clarify that: (1) all 1,247 cases in our dataset were obtained from publicly accessible sources where the arbitration proceedings and awards have been officially published or disclosed by the respective institutions; (2) no sealed or confidential case materials were accessed or incorporated; and (3) the term “confidentiality restrictions” in our previous draft referred to our inability to redistribute copyrighted full-text documents rather than any use of non-public materials.

Regarding full-text redistribution, we acknowledge that platforms such as italaw.com redistribute arbitration award texts under specific licensing arrangements with arbitration institutions. As an academic research project, we do not possess equivalent redistribution licenses for the complete corpus of award

texts. However, researchers can readily obtain all original source documents from the publicly accessible databases listed below, using the case identifiers we provide.

Publicly available arbitration case data can be accessed through the following official repositories: ICSID Cases Database (https://icsid.worldbank.org/cases/case-database), UNCITRAL Case Repository (https://uncitral.un.org), Permanent Court of Arbitration Case Repository (https://pca-cpa.org), Investment Treaty Arbitration Database (https://investmentpolicy.unctad.org/investment-dispute-settlement), and italaw Investment Treaty Arbitration (https://www.italaw.com).

To enable complete replication of our analysis, we provide Supplementary File 1, a comprehensive replication package containing all materials necessary for reproducing our results. This file includes: (1) Complete Case List providing case identifiers for all 1,247 cases (ICSID case numbers, ICC references, LCIA numbers, etc.), case names and parties, arbitration institution and procedural rules, award dates and outcome classifications, and direct URLs to publicly accessible award documents where available; (2) Extracted Feature Dataset containing all 127 engineered features in CSV format for each case, including 45 textual features (BERT-derived semantic scores), 64 numerical features (case characteristics, financial data), and 18 visual features (CNN-extracted representations), enabling researchers to reproduce our model training without re-processing original documents; (3) Complete Source Code with Python scripts for data preprocessing, feature extraction, model architecture implementation, training procedures, and evaluation metrics, including all library dependencies and version specifications; (4) Model Specifications documenting complete hyperparameter configurations, training procedures, and random seeds; and (5) Replication Guide with step-by-step instructions for obtaining original documents from public databases, reproducing our preprocessing pipeline, training the multimodal fusion model, and validating results against our reported performance metrics.

For researchers seeking to replicate our entire pipeline from original documents, we provide detailed data acquisition protocols specifying exactly which database queries and filters retrieve each case in our dataset. The feature extraction code processes standard arbitration award formats from the major institutions, enabling researchers to generate identical feature representations from the original documents. Model checkpoints (trained weights) are available upon reasonable request for academic research purposes, subject to completion of a data use agreement restricting use to non-commercial research. Researchers interested in collaboration or data access should contact the corresponding author at asd18103929689@163.com with specific details of intended use and institutional affiliation.

Authors' Contributions

Hao Wu conceptualized the research framework, designed the multimodal deep learning architecture, conducted the computational experiments, performed the

statistical analysis, and drafted the manuscript. Hao Wu also developed the data preprocessing pipeline, implemented the attention-based fusion mechanisms, and carried out the feature importance analysis and model validation procedures.

Jiajun Xu contributed to the theoretical framework development, participated in the literature review and background research, assisted with data collection and preprocessing, and provided critical review and revision of the manuscript. Jiajun Xu also contributed to the bilateral investment agreement analysis framework and supported the interpretation of legal and policy implications.

Both authors collaborated on the research design, methodology development, results interpretation, and manuscript preparation. All authors read and approved the final manuscript for publication.

References

[1] United Nations Conference on Trade and Development. (2024). Facts and figures on investor–State dispute settlement cases. IIA Issues Note, No. 3, 2024. https://unctad.org/publication/facts-and-figures-investor-state-dispute-settlement-cases

[2] United Nations Conference on Trade and Development. (2024). World Investment Report 2024: Investment facilitation and digital government. https://unctad.org/publication/world-investment-report-2024

[3] Katz, D. M., Bommarito, M. J., & Blackman, J. (2017). A general approach for predicting the behavior of the Supreme Court of the United States. PLOS ONE, 12(4), e0174698. https://doi.org/10.1371/journal.pone.0174698

[4] Cleary Gottlieb. (2024). Five international arbitration trends and topics for 2024. https://www.clearygottlieb.com/news-and-insights/publication-listing/five-international-arbitration-trends-and-topics-for-2024

[5] Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D., & Aletras, N. (2022). LexGLUE: A benchmark dataset for legal language understanding in English. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (pp. 4310-4330). https://doi.org/10.18653/v1/2022.acl-long.297

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30 (pp. 5998-6008).

[7] Medvedeva, M., Vols, M., & Wieling, M. (2020). Using machine learning to predict decisions of the European Court of Human Rights. Artificial Intelligence and Law, 28(2), 237-266. https://doi.org/10.1007/s10506-019-09255-y

[8] Transnational Matters. (2024). Bilateral vs multilateral investment treaties: Key contrasts. https://www.transnationalmatters.com/bilateral-vs-multilateral-treaties-bits-vs-mits/

[9] United Nations Conference on Trade and Development. (2024). International investment agreements navigator. https://investmentpolicy.unctad.org/international-investment-agreements

[10] Pinsentmasons. (2025). Major changes crystallising in investor-state dispute settlement. https://www.pinsentmasons.com/out-law/analysis/major-changes-investor-state-dispute-settlement

[11] American Bar Association. (2024). Using AI for predictive analytics in litigation. https://www.americanbar.org/groups/senior_lawyers/resources/voice-of-experience/2024-october/using-ai-for-predictive-analytics-in-litigation/

[12] Brookings Institution. (2024). A first look at outcomes under the No Surprises Act arbitration process. https://www.brookings.edu/articles/a-first-look-at-outcomes-under-the-no-surprises-act-arbitration-process/

[13] Enyo Law. (2025). ICC and LCIA arbitration statistics 2023: In-depth analysis and insights. https://enyolaw.com/news/icc-and-lcia-arbitration-statistics-2023-in-depth-analysis-and-insights/

[14] Chartered Institute of Arbitrators. (2024). Numbers don't lie: International commercial arbitration statistics 2023. https://www.ciarb.org/news-listing/numbers-don-t-lie-international-commercial-arbitration-statistics-2023/

[15] Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., & Lampos, V. (2016). Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. PeerJ Computer Science, 2, e93. https://doi.org/10.7717/peerj-cs.93

[16] Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., & Sun, M. (2020). How does NLP benefit legal system: A summary of legal artificial intelligence. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5218-5230). https://doi.org/10.18653/v1/2020.acl-main.466

[17] Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society, 4(2). https://doi.org/10.1177/2053951717743530

[18] Schreuer, C. (2009). The ICSID Convention: A Commentary (2nd ed.). Cambridge University Press.

[19] Smith, R. (2007). An overview of the Tesseract OCR engine. In Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) (Vol. 2, pp. 629-633). IEEE. https://doi.org/10.1109/ICDAR.2007.4376991

[20] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2980-2988). https://doi.org/10.1109/ICCV.2017.324

[21] Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423-443. https://doi.org/10.1109/TPAMI.2018.2798607

[22] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (pp. 4765-4774).

[23] Muthoo, A. (1999). Bargaining Theory with Applications. Cambridge University Press.

[24] Ashley, K. D. (2017). Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age. Cambridge University Press. https://doi.org/10.1017/9781316761458

[25] Surden, H. (2014). Machine learning and law. Washington Law Review, 89(1), 87-115.

[26] Alschner, W., & Skougarevskiy, D. (2016). Mapping the universe of international investment agreements. Journal of International Economic Law, 19(3), 561-588. https://doi.org/10.1093/jiel/jgw056

[27] Cui, Y., Jia, M., Lin, T. Y., Song, Y., & Belongie, S. (2019). Class-balanced loss based on effective number of samples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9268-9277). https://doi.org/10.1109/CVPR.2019.00949

[28] Engstrom, D. F., Gelbach, J., Ho, D. E., & Sharkey, C. M. (2020). Government by algorithm: Artificial intelligence in federal administrative agencies. Report for the Administrative Conference of the United States. https://www-cdn.law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf

[29] United Nations Conference on Trade and Development. (2024). World Investment Report 2024: Investment facilitation and digital government. https://unctad.org/publication/world-investment-report-2024

[30] Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11976-11986). https://doi.org/10.1109/CVPR52688.2022.01167

[31] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[32] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for

Computational Linguistics (pp. 4171-4186). https://doi.org/10.18653/v1/N19-1423

[33] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In 3rd International Conference on Learning Representations (ICLR 2015). https://arxiv.org/abs/1409.0473

[34] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In International Joint Conference on Artificial Intelligence (Vol. 14, No. 2, pp. 1137-1145).

[35] Prechelt, L. (2012). Early stopping - but when? In Neural Networks: Tricks of the Trade (pp. 53-67). Springer. https://doi.org/10.1007/978-3-642-35289-8_5

[36] Japkowicz, N., & Shah, M. (2011). Evaluating Learning Algorithms: A Classification Perspective. Cambridge University Press. https://doi.org/10.1017/CBO9780511921803

[37] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427-437. https://doi.org/10.1016/j.ipm.2009.03.002

[38] Kaufmann-Kohler, G., & Potestà, M. (2020). Can the Mauritius Convention serve as a model for the reform of investor-state arbitration in connection with the introduction of a permanent investment tribunal or an appeal mechanism? Analysis and Roadmap (3rd ed.). Geneva Center for International Dispute Settlement.

[39] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). https://doi.org/10.1145/2939672.2939778

[40] Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). https://christophm.github.io/interpretable-ml-book/

[41] Dolzer, R., & Schreuer, C. (2012). Principles of International Investment Law (2nd ed.). Oxford University Press.

[42] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. https://doi.org/10.1038/nature14539

[43] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

[44] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 2898-2904). https://doi.org/10.18653/v1/2020.findings-emnlp.261

[45] Franck, S. D. (2009). Development and outcomes of investment treaty arbitration. Harvard International Law Journal, 50(2), 435-489.

[46] Van Harten, G. (2012). Arbitrator behaviour in asymmetrical adjudication: An empirical study of investment treaty arbitration. Osgoode Hall Law Journal, 50(1), 211-268.

[47] Schultz, T., & Dupont, C. (2014). Investment arbitration: Promoting the rule of law or over-empowering investors? A quantitative empirical study. European Journal of International Law, 25(4), 1147-1168. https://doi.org/10.1093/ejil/chu075

[48] Limsopatham, N. (2021). Effectively leveraging BERT for legal document classification. In Proceedings of the Natural Legal Language Processing Workshop 2021 (pp. 210-216). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.nllp-1.22

[49] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778). https://doi.org/10.1109/CVPR.2016.90

[50] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations (ICLR 2015). https://arxiv.org/abs/1412.6980

[51] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2623-2631). https://doi.org/10.1145/3292500.3330701

[52] Egger, P., & Pfaffermayr, M. (2004). The impact of bilateral investment treaties on foreign direct investment. Journal of Comparative Economics, 32(4), 788-804. https://doi.org/10.1016/j.jce.2004.07.001

[53] Elkins, Z., Guzman, A. T., & Simmons, B. A. (2006). Competing for capital: The diffusion of bilateral investment treaties, 1960-2000. International Organization, 60(4), 811-846. https://doi.org/10.1017/S0020818306060279

[54] Yackee, J. W. (2008). Bilateral investment treaties, credible commitment, and the rule of (international) law: Do BITs promote foreign direct investment? Law & Society Review, 42(4), 805-832. https://doi.org/10.1111/j.1540-5893.2008.00359.x

[55] Strezhnev, A. (2018). Detecting bias in international investment arbitration. Journal of International Dispute Settlement, 9(3), 497-525. https://doi.org/10.1093/jnlids/idy020

[56] Puig, S., & Shaffer, G. (2018). Imperfect alternatives: Institutional choice and the reform of investment law. American Journal of International Law, 112(3), 361-409. https://doi.org/10.1017/ajil.2018.58

[57] Bonnitcha, J., Poulsen, L. N. S., & Waibel, M. (2017). The Political Economy of the Investment Treaty Regime. Oxford University Press. https://doi.org/10.1093/law/9780198719540.001.0001