IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 28 June 2025, accepted 28 July 2025, date of publication 5 August 2025, date of current version 13 August 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3596180

RESEARCH ARTICLE

Enhanced Hybrid Deep Learning Model With Improved Self-Attention Mechanism for Legal Judgment Prediction

G. SUKANYA AND J. PRIYADARSHINI
School of Computer Science and Engineering, Vellore Institute of Technology, Chennai Campus, Chennai 600127, India
Corresponding author: J. Priyadarshini (priyadarshini.j@vit.ac.in)

ABSTRACT The dawn of Natural Language Processing in the legal research field has achieved great heights. Legal Judgment Prediction (LJP) is one of the important tasks in legal intelligence that can assist attorneys and litigators in predicting judgments. Although current research techniques work effectively, they still have many drawbacks. Firstly, encoding lengthy case facts into vectors without losing information is one of the challenging tasks in LJP. Choosing an encoder that captures the syntax, semantics, and contexts of words can be extremely important for all downstream natural language tasks. Secondly, the features on which the deep learning model is trained also play an important role in testing the real-time cases. This research focuses on Indian cases that follow the common law, unlike civil law. Most of the existing LJP considers only civil law and does not emphasize the extraction of textual features. In this research a novel LJP approach has been suggested to overcome the issues raised above by improving the encoding part using the hybrid embedding method ELMo (Embeddings from Language Model) with Improved Principal Component Analysis (IPCA). Training a crucial set of features is done with a hybrid model, with a combination of Bi-GRU along with a modified attention mechanism and a deep-max-out network. The proposed Hybrid Deep Learning Model with Score Level Fusion (HDLMSF) is experimented with real-time Madras High Court criminal cases and compared with baseline classifier models. The results show that the proposed HDLMSF model has better prediction accuracy, 94.16% than other baseline classifiers.

INDEX TERMS Attention mechanism, BiGRU, deep max-out, feature extraction, ELMo word embedding, judgment prediction, principal component analysis.

I. INTRODUCTION

The primary function of the LJP for Indian Court cases is to predict judicial decisions based on the legal framework Bharatiya Nyaya Sanhita (BNS), which was earlier known as the Indian Penal Code (IPC) sections, by referring to the input case document, referred to as the fact description. Judges use BNS and identical jury verdicts in similar instances render a verdict on whether the case appeal is acquitted, adjourned, or dismissed. The legal justification for each litigation is complicated because of the numerous connections and links between earlier lawsuits [1]. Even in fundamentally identical cases, the complexity of the matter and the involvement of individuals often lead to differing judgments [2]. Also, most of the existing LJP works do not emphasize semantic feature extraction. References [3] and [4], which results in information loss at the initial stage itself. Significant advancements in NLP techniques have elevated legal research to a thriving and prominent field in the modern era [5]. Judgment prediction supports judges in decision-making, streamlines legal processes, and enhances case management efficiency. By minimizing the risk of misjudgments and reducing the likelihood of unjust rulings, LJP plays a crucial role in advancing the overall legal framework of society [6]. Existing LJP works experimented with different datasets of legal cases in different languages, like the Chinese Judgment

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wei.

Network [7], the CAIL dataset on Chinese civil cases [8] [9], [10]. Some of the LJP works were in European [11] language [12], French [13], and Swiss [14] cases. Legal judgments typically comprise several components, such as predicting law sections [15], charge type [16], penalty terms, and final judgment. The mounting backlog of cases and the chronic shortage of judges challenge the Indian judiciary. In the High Courts, the least burdened judges have around 15-16 minutes to hear a case, while the busiest manage with only about 2.5 minutes per case [52]. A study from DAKSH, an NGO, analyzes that it takes around 384 days from evidence to final judgment in Indian Subordinate courts [17], [18]. One of the solutions is to add digital assistance to the case workload to speed up the judgment. In this investigation, criminal cases are used, as there is no systematic procedure to solve the diverse nature of civil cases [19]. Also, earlier works predict the case based on the law article sections alone, leaving the embedded semantics present in the fact description [20]. The judgment prediction in this research work focuses on whether the case is acquitted or dismissed. Earlier research on LJP used static word embeddings, like Word2Vec or GloVe, which fail to capture context-specific meanings, leading to inaccuracies in understanding polysemous words. Additionally, they cannot adapt to domain-specific details, limiting their effectiveness in specialized fields like law. Some challenges, like sparse data distribution, arise in word embedding when embedding is done using high-dimensional spaces. The more dimensions, the higher the risk of overfitting, as the model may focus on irrelevant features or noise in the data. Reducing dimensions while retaining semantic information is crucial to overcoming these challenges [21].

An attention mechanism in a deep learning model enables the classifier to focus on the most relevant parts of the input when making predictions, improving understanding in tasks with complex or lengthy data. LJP models that are based on TF-IDF [22], static embedding, and RNN without using an attention mechanism treat all input elements equally, often dilute focus, and miss important context, especially in lengthy or complex data. On the other hand, transformer-based models such as BERT, LegalBERT, and DL models with self-attention and hierarchical attention [23] leverage attention mechanisms to enhance the interpretability and accuracy of legal decision-making, ensuring that context and relevant facts are prioritized during prediction. While self-attention focuses on relationships between words, it may fail to capture the intricate domain-specific legal context fully. Legal texts often require a specialized understanding of relationships between legal terms, precedents, and case-specific facts, which self-attention may overlook without additional fine-tuning or domain adaptation. These challenges can lead to inefficiencies and limitations in accurately predicting legal judgments based on case facts.

This research work mainly concentrates on improving self-attention mechanisms and was carried out on Madras High Court criminal cases in the English language and the main contributions of this paper are presented as follows:

- Proposed an enhanced feature representation by providing an ensemble word embedding with ELMo and USE (Universal Sentence Encoder) and reducing the dimensions using Improved Principal Component Analysis (IPCA) for more efficient and informative input features.
- Introduced a Bi-GRU model with an improved Self-attention mechanism to enhance the capability of the architecture model to predict more accurately, where the attention layer is integrated within the Bi-GRU to extract more complex and relevant features.
- Developed a modified Score Level Fusion(SLF) mechanism as an enhancement over the conventional SLF, enabling better prediction accuracy for legal cases. This improvement is achieved by updating the traditional SLF model using a cubic mapping approach.

The rest of the paper is organized as follows: literature review in Section II, methodology of the prediction model in Section III, results are discussed in Section IV, and the work is concluded in Section V.

II. RELATED WORKS

Legal NLP has excelled in automating intricate text analysis, particularly within the context of Indian jurisdiction. This section describes the relevant literature on the prediction of legal judgments in the context of deep learning models with and without an attention mechanism and the influence of different word embedding methods.

A. DEEP LEARNING MODELS WITH ATTENTION MECHANISM

In 2023, Chen et al. [24] developed the Mulan model, which uses a multiple residual article-wise attention network that focuses on relevant law articles for each case. It integrates residual connections to capture dependencies between different legal provisions. The model was better able to identify relevant legal provisions thanks to the article-wise attention mechanism, which produced more correct judgment predictions. Deeper network training was made possible by residual connections without the problems of gradient vanishing. Addressing the interdependencies among various law articles and ensuring the model's scalability to large legal corpora were primary challenges. In 2022, Bertalan and Ruiz [25] employed Hierarchical Attention Networks (HAN) to predict judicial outcomes in the Brazilian legal system, capturing the hierarchical structure of legal documents, assigning attention weights at both word and sentence levels to identify the most important information. The attention mechanism enhanced interpretability by highlighting influential words and sentences, aiding legal professionals in understanding the basis of predictions. The complexity of legal language and the need for extensive labeled data posed significant challenges. Additionally, ensuring the model's

generalizability across different legal domains required careful consideration. In 2024, Nigam et al. [26] compared transformer-based models, including InLegalBERT, BERT, and XLNet, alongside large language models (LLMs) like Llama-2 and GPT-3.5 Turbo, to predict judgments in realistic scenarios using only information available at the time of decision, such as case facts, statutes, precedents, and arguments. It simulates real-world conditions by focusing on the information available when a case is presented in court, avoiding retrospective analyses. Despite advancements, the study finds that LLMs have not yet achieved expert-level performance in judgment prediction and explanation tasks, indicating room for improvement. In 2020 Xu et al. [16] proposed the LADAN model, which combines a graph neural network with an attention mechanism to differentiate between similar confusing law articles. The integrated attention mechanism allowed the model to focus on critical distinctions between law articles, reducing misclassification rates in legal judgment prediction. Two major challenges were handling the complexity of graph-based representations and distinguishing law articles with overlapping applicability. Existing work features and challenges of deep learning models with attention mechanism are provided in Table (1) In 2021, Kong et al. [27] introduced a hierarchical version of BERT to predict legal judgments in English texts. The model addresses BERT's length limitations by hierarchically processing long legal documents. It also offered insights into the decision-making process through attention weights. Handling lengthy legal documents and capturing long-range dependencies within the text were primary challenges.

TABLE 1. Challenges in existing judgment prediction work with attention mechanism.

Author [citation]	Method	Features	Challenges
Huu et.al [28]	EMNLNM	Integrated attention-based Bi-LSTM memory and dilated skip residual CNN. Captures both sequential and hierarchical features in legal texts.	Complex dependencies in capturing both local and global contextual information.
Chen et.al [29]	CMLEA (Cross-Attention Mechanism and Label-Enhancement Algorithm)	Constructs a crime similarity graph and utilizes discriminative keywords for each charge, Integrates them into case embeddings to capture deep semantic representations.	Ensuring consistency across multiple subtasks in LJP is a challenge
Shen et.al [30]	Legal event type with Pedal attention mechanism	Identifies key event information within fact descriptions. Combines event-aware and event-free representations of case fact.	Balancing event-aware and event-free representations is complex
Bao et.al [31]	LegalAtt	Focuses on relevant parts of fact descriptions corresponding to specific charges.	Aligning fact descriptions with appropriate legal charges; managing the variability in legal language and case specifics.
Shelar et.al [32]	a deep BiLSTM classifier (TWO-BiLSTM) model based on Texas wolf optimization	Attention mechanism to effectively capture intricate linguistic patterns within legal documents is used. Feature extraction is enhanced through statistical methods and Principal Component Analysis (PCA)	Optimizing the BiLSTM model's hyperparameters to accurately identify complex patterns within legal documents pose challenge

B. DEEP LEARNING MODELS WITHOUT ATTENTION MECHANISM

In 2024, Dong et al. [33] proposed a graph-based model that utilizes contrastive learning with data augmentation to enhance LJP. The model constructs fact graphs from legal documents and applies supervised contrastive learning to capture semantic representations without relying on attention mechanisms. It can be difficult to create and enhance fact graphs from complex legal documents efficiently. The graph structures and contrastive learning show enhanced performance in LJP tasks, offering a new method that does not require attention processes. Xu et al. [34] introduce the D-LADAN model, which constructs a graph among law articles based on their text definitions and employs a graph distillation operation to distinguish articles with high prior semantic similarity. It also presents a momentum-updated memory mechanism to dynamically sense posterior similarity between law articles, addressing data imbalance issues. Managing the complexity of graph-based representations presents a significant difficulty, as does addressing confusion between legal articles caused by both past semantic similarity and data imbalance.

Abbara et al. [22] developed a system utilizing deep learning models, including Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM), for predicting judgment outcomes from Arabic case scripts, specifically in marriage cases. Handling the complexity of Arabic legal texts, achieving high prediction accuracy without the use of attention mechanisms poses a significant challenge. The Support Vector Machine (SVM) model with word2vec and Logistic Regression (LR) with TF-IDF achieve high accuracy rates, demonstrating the effectiveness of traditional machine learning models combined with deep learning techniques in LJP tasks.

TABLE 2. Challenges in existing judgment prediction work without attention mechanism.

Author [citation]	Method	Features	Challenges
Zhang et.al [37]	LJPCheck	23 Functionality test done with models taken for comparison.	Developing comprehensive functional tests that accurately evaluate different aspects of LJP models.
Haidar et al [38]	Linear Regression, Decision Trees, and Random Forests,	Random Forest algorithm achieved the highest accuracy of 91.05%	lack of sufficient labeled data
Daniyal et. al [39]	LSTM+CNN	Captures both sequential dependencies and spatial features within legal documents. Optimal feature selection is used.	Complexity of Legal language and data preprocessing pose a challenge
Shang et.al [40]	CNN+PCA	Integrates Convolutional Neural Network (CNN) for text feature extraction and with PCA for data feature dimensionality reduction. Employs a genetic algorithm for parameter optimization	Modeling dependencies in legal data and optimization of parameters is complex

Guo et al. [35] have conducted a study that constructs detailed graphs representing various elements of legal cases, such as facts, charges, and applicable laws. The model incorporates external legal knowledge bases, including statutes and precedents, to enrich the contextual understanding of cases. Developing fine-grained element graphs requires meticulous extraction and representation of legal case components, posing challenges in accurately modeling the complexities of legal texts. Zhang et al. [36] proposed a contrastive learning approach where they employed a contrastive loss function to distinguish between similar and dissimilar cases, enhancing the model's ability to generalize across different legal domains. Existing work features and challenges for LJP works using classifier models without attention mechanisms are provided in Table (3).

Although a lot of research has been done on using deep learning models to predict court rulings, attention must be paid to word embedding in the early stages and the features that are utilized to train deep learning classifiers. The research work attempts to advance the state-of-the-art in legal judgment prediction by incorporating improved word embedding and a hybrid deep learning model with a modified attention mechanism, providing a thorough framework that tackles the particular difficulties, thereby minimizing the loss.

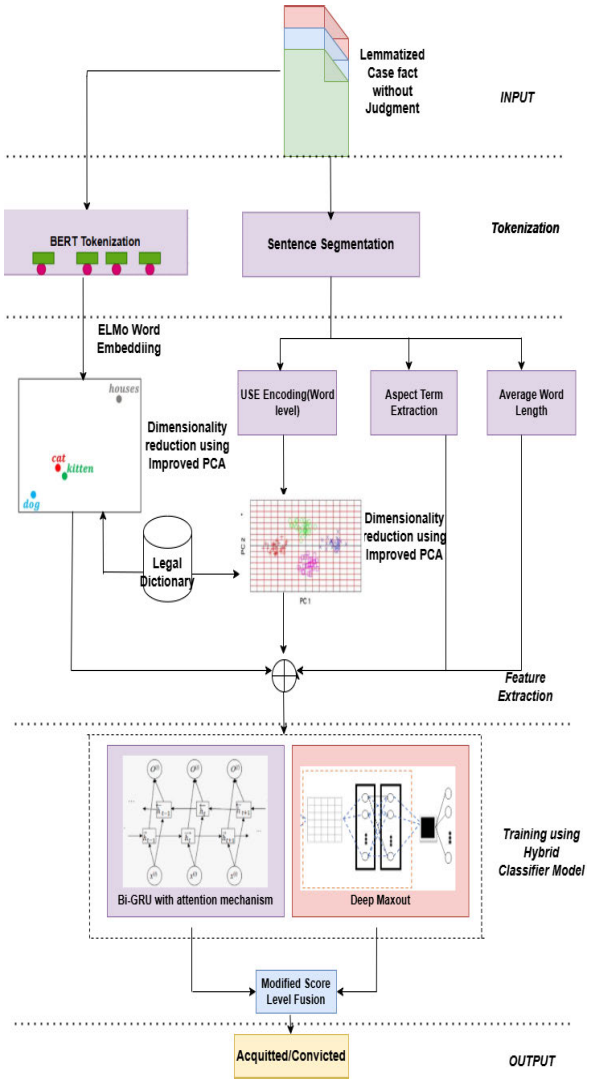

FIGURE 1. Overall structure of judgment prediction model using HDLMSF.

III. PROPOSED METHODOLOGY

This section provides a thorough explanation of our HDLMSF framework. The architecture is made to tackle the particular difficulties of legal judgment prediction (LJP) using a multi-phase strategy that combines the strength of embedding techniques with legal dictionaries and hybrid deep learning models with modified score-level fusion. Figure (1) shows the HDLSMF framework of the research article in general.

A. PREPROCESSING OF INPUT TEXT VIA BERT TOKENIZATION

In the preprocessing stage of the proposed judgment prediction method, case facts without judgment are taken into consideration as input. BERT tokenizer [41] is used in this phase after stopwords have been removed. There are three

common types of tokenization: Sentencepiece tokenization, Wordpiece tokenization, and Byte-pair encoding. To generate word-level tokens from the case text T into T_p, a wordpiece tokenizer is utilized. Word embedding models such as Universal Sentence Encoding (USE) and ELMo are used to encode the word-level tokens. Combining these embeddings allows incorporation of both context-sensitive (ELMo) and global/static (USE) knowledge, leading to richer and more nuanced word representations reducing information loss to a great extent.

B. FEATURE EXTRACTION

Character-level convolutions (CharCNNs) are typically used by ELMo, which enables it to efficiently handle unusual and out-of-vocabulary (OOV) words. More accurate representations of domain-specific terminology and morphologically rich languages. However, WordPiece, which is used in BERT, RoBERTa, and GPT, separates uncommon or unknown words into subword units, which may lose their morphological or semantic structure and are unable to properly capture internal word structures or spelling variants. It also uses a set vocabulary of subwords. Subwords must be used to approximate any words that are not in this vocabulary. However, ELMo is vocabulary independent because it creates word embeddings dynamically. In contrast to BPE/Wordpiece, which segment words according to frequency, ELMo employs entire word embeddings constructed from characters, guaranteeing consistent token borders and preventing tokenization-induced ambiguity. Even if the word is absent from the training set, it still functions. The following example shows the comparison of embeddings for an unknown word, ''Unconstitutionality''.

Tokenization Comparison Example: The following illustrates how the word *''Unconstitutionality''* is tokenized differently across models:

- **BERT Tokens:** `['unconstitutional', '##ity']`
- **RoBERTa Tokens:** `['un', 'const', 'itution', 'ality']`
- **GPT Tokens:** `['unconstitution', 'ality']`
- **ELMo Tokens:** `['unconstitutionality']`

ELMo's thorough contextualization from both sides at all internal layers is another important benefit that allows it to catch subtle syntactic and semantic connections, even in regimes with minimal data. ELMo computes word embeddings as a learnt weighted combination of all intermediate layers from its biLSTM architecture, in contrast to BERT and RoBERTa, which only employ the final layer representation for downstream tasks (unless fine-tuned). Because of this, it works especially well for domain adaptation in specialized fields like legal natural language processing, where task-specific subtleties could not be well covered by pretraining corpora of huge LLMs. The features to be extracted from the preprocessed text, T_p via BERT tokenization for the research are elaborated as follows:

1) ENSEMBLE WORD EMBEDDING

Ensemble word embeddings combine multiple embeddings from different models or techniques, offering distinct advantages over single word embeddings. Word-level tokens of the case facts from the BERT tokenizer are initially encoded using contextualized Embeddings from Language Models(ELMo) [42]. Conventional ELMo utilizes bidirectional LSTMs, i.e., it concatenates both the forward and backward contexts concurrently. Although ELMo uses bi-directional LSTMs to capture context, LSTMs struggle with capturing long-range dependencies in text, resulting in some information loss [19]. To overcome this limitation, the ELMo word embedding feature introduced in this research work is employed with USE. Using Universal Sentence Encoder (USE) as a word embedding alongside ELMo can complement their respective strengths, providing a significant benefit. ELMo ensures a fine-grained, word-level understanding of contextual representations. USE for word-level encoding makes sense for scenarios prioritizing efficiency, scalability, and semantic representation [10]. Since both ELMo and Universal Sentence Encoder (USE) generate high-dimensional embeddings, to enhance representation quality, ELMo and Universal Sentence Encoder (USE) embeddings are combined by concatenating their vectors, resulting in high-dimensional feature representations of 1536 dimensions (ELMo: 1024, USE: 512). To address the computational inefficiency and sparsity issues associated with such high-dimensional data, Improved Principal Component Analysis (IPCA) is employed, which uses a winsorized mean to reduce sensitivity to outliers. This approach produced robust and compact representations, reducing the ELMo and USE embeddings to 30 dimensions each, thereby yielding a combined 60-dimensional vector. The IPCA transformation preserved over 95% of the original variance while significantly enhancing computational efficiency and generalization capability In high-dimensional spaces, data points tend to become sparse and equidistant. This can make it difficult for similarity algorithms to distinguish between relevant and irrelevant relationships, leading to degraded performance in certain tasks. Also, limited data can lead to overfitting, as the model may focus on irrelevant patterns in the data rather than generalizable features. So, Dimensionality reduction techniques like PCA can be used to mitigate this issue. PCA is the most popular technique used in dimensionality reduction. By this technique, high-dimensional text is transformed into low-dimensional text while maintaining the variance of the preprocessed text at maximum values.

The primary drawback of conventional PCA in the context of the mean is its reliance on the global mean of the dataset. This assumption can lead to challenges in certain scenarios, like outliers. The conventional PCA [43] is updated as improved PCA as in Eq. (1) using the winsorized mean instead of the mean. Because features that drastically differ from the rest of the data, or outliers, damage the PCA model. Here, the winsorized mean X_b is taken into account in the research task by substituting the extreme values because of

its particular capacity, such as the values set to the lowest or maximum permitted within a given range. In addition, the winsorized rules preserve the principal components by diminishing the influence of the outliers on the covariance matrix. Also, it helps to retain the relationship between the attributes while minimizing the distortion due to outliers.

$$Y_{ab} = (X_{ab} - \bar{X_b}) / \sum X_b \tag{1}$$

$$\bar{X_b} = 1/n((c+1)X_{c+1}) + \sum_{a=c+1}^{n-c-1} + (c-1)X_{n-c} \tag{2}$$

Eq. (3) represents the ensemble word embedding feature, which is a fusion of ELMo word-embedded sequence and USE word-embedded sequence using the weighted average method. In the weighted average method, each feature is formulated by multiplying by its weighting factor designed using a legal dictionary and then sum them.

$$I_{ELMoFeat} = w_1 X_1 + w_2 X_2 \tag{3}$$

Here, the values of the weights w_1 and w_2 are assigned as 0.7 and 0.3, respectively, and the ELMo embedding sequence and USE embedding sequence are represented as x1 and x2, correspondingly. Thus, an improved ensemble word embedding feature is extracted, and it is indicated as $I_{ELMoFeat}$.

2) ASPECT TERM EXTRACTION

The term 'aspect term' denotes a specific attribute or feature of a service or product in judgmental facts. It can either be an explicit term or an implicit term in a sentence. The aspect terms or targets in a sentence are identified by an aspect term extraction mechanism [44]. The ATE feature is extracted from the preprocessed text, T_P, as ATE_{Feat}. The following figure 2 shows an example of aspect term extraction of criminal case sentences.

"The stolen vehicle was recovered near the suspect's home."	*"The victim sustained multiple injuries due to a blunt object."*
• Extracted Aspect Terms: ◦ "stolen vehicle" (evidence) ◦ "suspect's home" (location)	• Extracted Aspect Terms: ◦ *"victim" (entity)* ◦ *"multiple injuries" (evidence)* ◦ *"blunt object" (weapon)*

FIGURE 2. Criminal case sentences showing aspect term extraction.

3) AVERAGE WORD LENGTH

The evaluation of the average length of words in a text or a set of texts is called the average word length. The resultant value obtained from this evaluation is in numerical form, which indicates the average number of characters per word in a given text. By dividing the overall word length by the text's word count, the average word length is calculated using this evaluation. Therefore, the extraction of average word length is performed on preprocessed text, T_P, and it is indicated as AWL_{Feat}. Finally, the extracted ensembled word embedding features, ATE and average word length are summed as a whole (Eq. (4)).

$$T_{Feat} = I_{ELMoFeat} + ATE_{Feat} + AWL_{Feat} \tag{4}$$

C. JUDGMENT PREDICTION VIA HYBRID DL APPROACH

The hybridization of two DL approaches implements the proposed approach on judgment prediction, proposed Bi-GRU with improved self-attention mechanism, and Deep maxout algorithms. A modified score-level fusion is used to determine the prediction results. This phase is broadly explained in the sections below. The diagrammatic representation of the proposed Hybrid Deep Learning Model with Score Level Fusion (HDLMSF) model is illustrated in Figure 3.

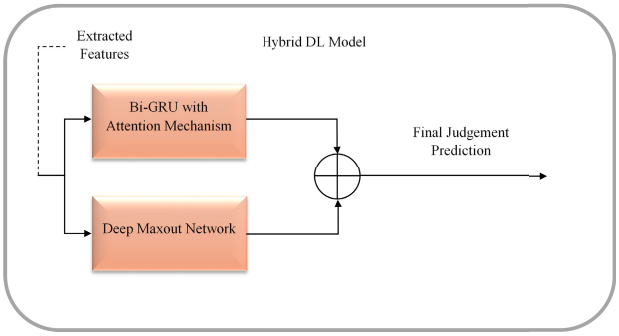

FIGURE 3. Hybridized Bi-GRU with modified attention mechanism and deep maxout network.

1) PROPOSED BI-GRU WITH IMPROVED SELF-ATTENTION MECHANISM

A Bi-GRU classifier [53] is a sequential processing model that utilizes two Gated Recurrent Units (GRUs). One GRU processes input data in the forward direction, while the other processes it in the backward direction. The outputs from both GRUs are combined at the same output layer, enabling a more comprehensive representation of the data. This classifier enhances the traditional RNN architecture by introducing input and forget gates, effectively addressing gradient explosion and vanishing gradients. In the proposed HDLMSF model, Bi-GRU is employed for its gated structure, which is well-suited for preserving contextual information. [5]. In standard self-attention, each word gets a weight (α_i) that tells the model how important it is in the sentence context. In this proposed model, **HDLMSF**, self-attention is enhanced by integrating a *reset gate* (from GRU logic) into the attention mechanism. This helps:

- Dynamically control how past information (memory) influences the current word representation.
- Filter out irrelevant past features by adjusting the contribution of previous steps using the reset gate (r_t).
- Combine attention weights with gated recurrence, making attention more context- and time-aware.

So, instead of using attention weights alone, the model transforms features using a reset-gated mechanism before computing the attention-weighted sum.

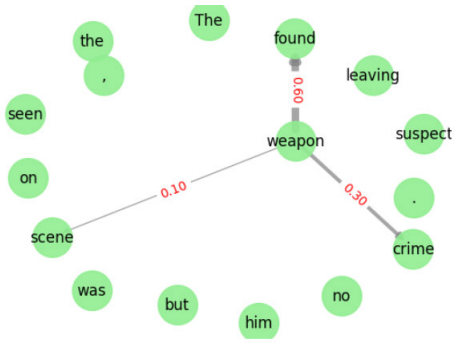

FIGURE 4. Self attention visualization.

2) IMPROVED SELF-ATTENTION MECHANISM

Attention mechanisms capture long-range dependencies and contextual relationships better than traditional architectures like RNNs or LSTMs by focusing on important parts of the input data, reducing the burden of processing irrelevant information. The self-attention mechanism, also known as scaled dot-product attention, is a key concept in modern deep-learning architectures, especially in Transformer models. For each input element in a sequence, the self-attention mechanism computes how much attention it should pay to other elements in the sequence. In the proposed hybrid model, an improved self-attention mechanism layer is added along with BI-GRU. It enables the model to focus on certain parts of the input by allotting various weights to various parts of the input. This weighting factor is assigned based on the relation between the input and the task performed by the model. The Bi-GRU network is constructed with one forward and backward GRU layer, an Improved self-attention mechanism layer, and a fully connected layer. The extracted feature set, T_{Feat}, is given as an input to the judgment prediction phase. Figure 4 shown below depicts the self-attention scores of the criminal case sentence. ' The suspect was seen leaving the crime scene, but no weapon was found on him. Attention mechanisms capture long-range dependencies and contextual relationships better than traditional architectures like RNNs or LSTMs by focusing on important parts of the input data, reducing the burden of processing irrelevant information. The self-attention mechanism, also known as scaled dot-product attention, is a key concept in modern deep-learning architectures, especially in Transformer models. For each input element in a sequence, the self-attention mechanism computes how much attention it should pay to other elements in the sequence. In the proposed hybrid model, an improved self-attention mechanism layer is added along with BI-GRU. It enables the model to focus on certain parts of the input by allotting various weights to various parts of the input. This weighting factor is assigned based on the relation between the input and the task performed by the model. The Bi-GRU network is constructed with one forward and backward GRU layer, an Improved self-attention mechanism layer, and a fully connected layer. The extracted feature set, T_{Feat}, is given as an input to the judgment prediction phase. Figure 4 shown below depicts the self-attention scores of the criminal case sentence,'

The suspect was seen leaving the crime scene, but no weapon was found on him.

$$\overrightarrow{f_t} = \overrightarrow{GRU}(w_t) \tag{5}$$

$$\overleftarrow{f_t} = \overleftarrow{GRU}(w_t) \tag{6}$$

$$f_t = [\overrightarrow{f_t}, \overleftarrow{f_t}] \tag{7}$$

$$x_t = tanh(w_t.f_t + B_v) \tag{8}$$

$$\alpha_t = e^{(w_v \cdot x_t^Q)} / \sum_{t=1}^{n} e^{(w_v \cdot x_t^Q)} \tag{9}$$

$$X_t' = \sum_{t=1}^{n} \alpha_t \cdot f_t \tag{10}$$

where

α_t = Attention weight of the word vector
w_v = Weight parameter of conventional Bi-GRU
B_v = Bias parameter of conventional Bi-GRU
tanh = Activation Function

The above (9) expresses the mathematical equation for the weight of the word vector attention, and (10) represents the mathematical expression for the word vector, X_t. The proposed Bi-GRU with an improved attention mechanism is formulated in (11) and (12) instead of (8) as the standard self-attention mechanism suffers from an inherent locality bias. The reset gate information for BiGRU has been added to overcome the above limitation.

$$f_t = tanh(w_a.[r_t \times f_t, x_t] + B_v) \tag{11}$$

$$r_t = \sigma(w_v \cdot [f_t, x_t] + B_v) \tag{12}$$

(11) and (12) are improved by utilizing reset gate. Where,

w_a and B_a = weight and bias parameters of the proposed Bi-GRU with improved Self-attention mechanism
σ = sigmoid function
$\cdot$ = Dot product
x_t = Input vector of time t
r_t = Reset gate of time t
f_t = Hidden state vector or output vector of time t

In proposed form, the output of hidden states of forward and backward GRU units are formulated in (13) and (14).

$$\overrightarrow{f_t} = \overrightarrow{GRU}(x_t, \overrightarrow{f_t}) \tag{13}$$

$$\overleftarrow{f_t} = \overleftarrow{GRU}(x_t, \overleftarrow{f_t}) \tag{14}$$

$$f_t = [\overrightarrow{f_t}, \overleftarrow{f_t}] \tag{15}$$

Consequently, the corresponding input feature set, x_t is evaluated using information gain to select useful information from x_t. Information gain [13] is a technique that is used for reducing the dimensions of the dataset by removing irrelevant features. Therefore, the required set of features is obtained as in (16) using information gain. The general equation for

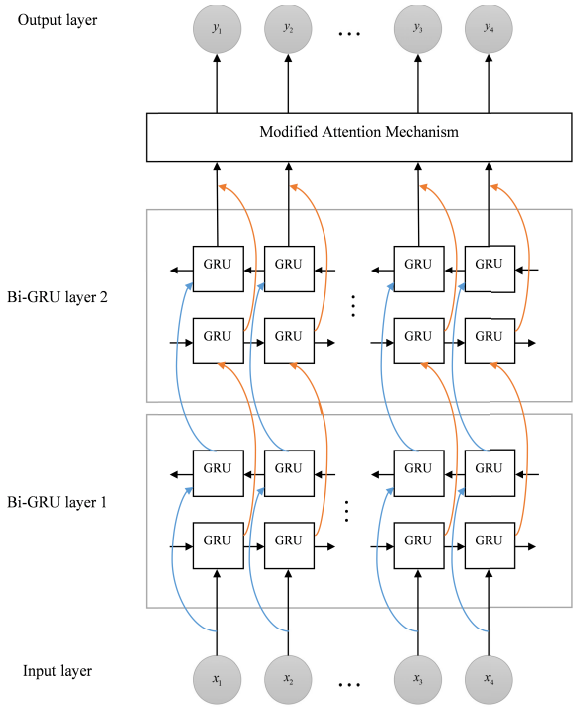

FIGURE 5. Architecture of Bi-GRU with Improved Self-attention.

information gain is shown in (17).

$$x_t = x_t + IFG \tag{16}$$

$$IFG = H(S) - H(S/j) \tag{17}$$

where, $H(S)$- Entropy of x_t, $H(S/j)$- Conditional entropy for the dataset given the variable j. Figure 5 shows the architecture of BiGRU with the Improvised Self-attention score. The intermediate prediction results obtained from the input, T_{Feat}, in the proposed Bi-GRU with improved attention mechanism are represented as BAM_{out}.

One of the important features in the Bi-GRU architecture is the reset gate, which is essential for regulating how much historical data is lost at each time step. This reset gate allows the GRU to ignore irrelevant earlier context and concentrate on more immediate signals by allowing the model to 'reset' its memory. By enabling the model to flexibly adapt its memory to longer or shorter dependencies as needed, this is especially helpful in reducing locality bias, which is the tendency of sequence models to overprioritize recent or neighboring tokens.

3) DEEP MAXOUT NETWORK

The Deep Maxout Network [45] is a widely used deep learning approach known for its faster convergence compared to other methods. The maxout unit in the deep maxout network typically employs ReLU and leaky ReLU activation functions. Additionally, the use of a piecewise linear function, combined with dropout, enhances the network's robustness by increasing the diversity of inputs. Furthermore, the proposed HDLMSF model benefits from the activation function of the deep maxout network, which improves its overall performance. The extracted features, such as improved ensemble word embeddings, aspect term extraction, and average word length, are fed into the deep maxout network. The mathematical formulation of the deep maxout network is provided in (18).

$$q(x) = max(R_{uv}) \text{where } v \in [1, \eta] \tag{18}$$

$$R_{uv} = x \cdot w_{uv} + B_{uv} \tag{19}$$

After the training process is done in the deep-maxout classifier model using T_Feat, the resultant intermediate judgement prediction results are represented by DMN_{out}. Thereby, the intermediate prediction results obtained from Bi-GRU with modified attention mechanism and deep maxout network are given to modify the score level fusion process to obtain final prediction result on judgment. Figure 6 illustrates the architecture of deep maxout classifier.

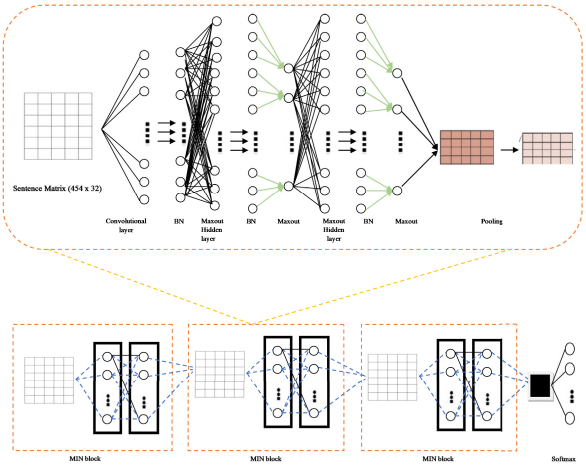

FIGURE 6. Architecture of deep Maxout classifier.

4) MODIFIED SCORE LEVEL FUSION

The hybrid deep learning model, which consists of a proposed Bi-GRU classifier with an improved self-attention mechanism and Deep Maxout trained using TFeat, which consists of an ensemble word embedding feature, Aspect Term Extraction, and average word length extracted from the word tokens of case acts without judgment. Score-level fusion is a technique used in hybrid deep learning models to combine scores generated by multiple classifiers or models. This process improves decision-making by aggregating information from various sources. The general formula used for Min-Max normalized score-level fusion is shown in (20).

$$S_a' = (S_a - min(S_a))/max(S_a) - min(S_a) \tag{20}$$

$$F_{slf} = \sum_{a}^{z} (S_{BAMOUT_a} + S_{DMNOUT_a}) \tag{21}$$

where S_a: Raw score from the a-th model.

$min(S_a)$: Minimum score of the a-th model's score range.

$max(S_a)$: Maximum score of the a-th model's score range.

$S_a{}'$: Normalized score from the a-th model, rescaled to the range [0, 1].

The normalized fusion model, while widely used for score-level fusion in hybrid systems, has several potential drawbacks, such as sensitivity to outliers in the score range and loss of interpretability, as it may no longer reflect the original confidence or probability meaningfully. The above limitations of the normalized fusion approach can be addressed, at least partially, by employing weights generated using a cubic map. Using weights derived from a cubic mapping function introduces a non-linear transformation that can capture intricate relationships between models and their respective contributions to the fusion process [26]. The modified SLF is updated by a cubic map, which is expressed in (22).

$$S_a{}' = (S_a - median)/MAD \tag{22}$$

where

MAD = $mean(|S_a - mean|)$ and the updated fusion equation of (21) is expressed in (23).

$$F_{slf} = [w_r/S_{BAMout_a} + w_m/S_{DMNout_a}]/w_r + w_n \tag{23}$$

where

$$w_r = 2.59 \cdot c_k \cdot (1 - c_k^2)$$

and w_m = 1-w_r

S_{BAMout_a}- Predicted intermediate result of Bi-GRU with modified attention mechanism

S_{DMNout_a} - Predicted intermediate result of deep maxout network.

Min-Max normalization in the traditional formula is replaced with Median Absolute Deviation (MAD) normalization. This makes the fusion more robust to skewed or noisy distributions, as the median and MAD are less sensitive to outliers than min/max or mean. The weights w_r: Weight generated for Bi-GRU output via cubic mapping and w_m: Weight generated for Deep Maxout output via cubic mapping pose significant importance. The cubic term boosts confident predictions. It suppresses uncertain predictions, enhancing the signal-to-noise ratio. Unlike linear fusion, this makes the final decision non-linearly sensitive to high-confidence scores from more reliable models

From the final score, the judgment for legal cases, whether admitted or convicted, is predicted.

D. RESULTS AND DISCUSSION

1) DATASET DESCRIPTION

This research utilizes a dataset sourced from Manupatra, an online Indian legal platform that hosts a comprehensive collection of case documents. The dataset comprises approximately 500 completed criminal case documents from the Madras High Court, with judgments categorized as either dismissed or allowed. The dataset encompasses 15 distinct types of raw criminal cases, which have been web-scraped and processed into necessary labels, including case notes, facts, judgments, applicable legal sections, and judgment labels (allowed or dismissed). This structured data is organized in a .csv format, generated with the assistance of entity extraction techniques and regular expressions. The performance of the HDLMSF-based judgment prediction was compared with traditional techniques, such as SVM [7] LSTM+CNN [46], NN, KNN, GRU, Bi-GRU, LSTM, and Deep Maxout. Moreover, it was examined with respect to sensitivity, FNR, MCC, accuracy, and other metrics. Legal datasets are often hard to obtain, especially with consistent labels, metadata, and judgment text. It generalizes well for a proposed model MHAN which was experimented with in both Madras High Court criminal cases and Supreme Court civil cases of the Indian legal system [23]. Future work will explore cross-jurisdictional generalizability by experimenting on diverse datasets from other Indian Courts.

2) ASSESSMENT ON HDLMSF AND CONVENTIONAL STRATEGIES REGARDING POSITIVE METRIC FOR JUDGMENT PREDICTION

Figure 7, 8, 9, 10 explains the positive metric evaluation on HDLMSF compared to the SVM, LSTM+CNN, NN, GRU, Bi-GRU, LSTM, and Deep Maxout for the prediction of judgment. For the accurate prediction of judgment, the model should acquire greater positive metric ratings. Positive metrics refer to measures derived from correctly identifying the positive class (True Positives, TP). These metrics are particularly important in scenarios where detecting positive instances is critical. Sensitivity measures the proportion of actual positives correctly identified by the model. A low sensitivity indicates that many positives are missed. Precision measures the proportion of predicted positives that are actual positives. In datasets where the positive class is rare, metrics like Accuracy can be misleading, and positive metrics (e.g., recall, precision) give a better picture of model performance. In the proposed method, HDLMSF, the accuracy is 0.9586 for a training percentage of 80, though the traditional schemes acquired minimal accuracy ratings. Simultaneously, while contrasting the HDLMSF over previous strategies, the sensitivity and specificity rates are much better for the HDLMSF technique. On examining the findings of the HDLMSF, a training rate of 70% achieved the highest specificity of 92.2039 compared to other values. Moreover, the evaluation is done with regard to the sensitivity metric by adjusting the training rate to 60, 70, 80, and 90. In those training rates, the HDLMSF accomplished the highest sensitivity of 92.5558, 95.2381, 96.1772, and 97.1446, respectively. With better predicted outputs, the evaluation demonstrated that the HDLMSF method had improved efficacy for judgment prediction. The improvement in accuracy of the HDLMSF model is due to the use of an ensembled word embedding-based feature process with the hybrid model (Bi-GRU and Deep Maxout).

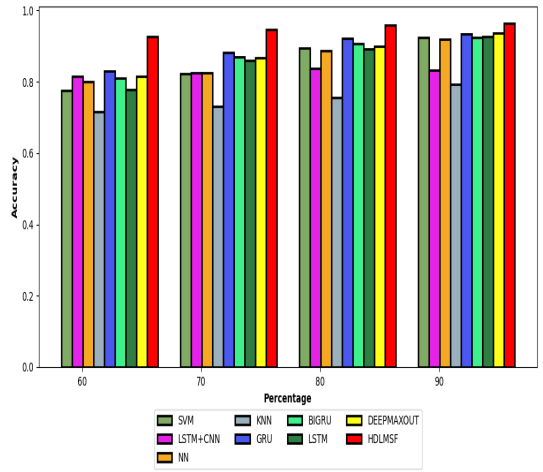

FIGURE 7. Accuracy.

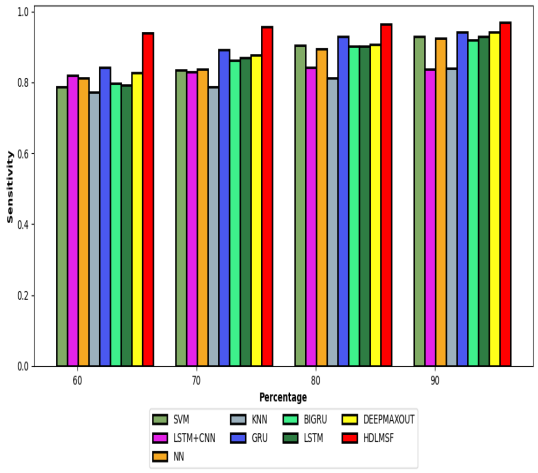

FIGURE 9. Specificity.

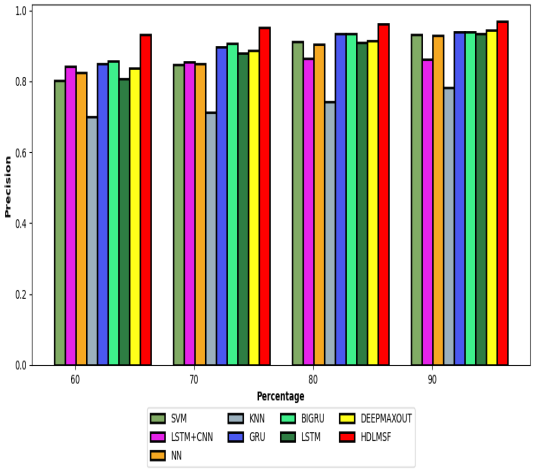

FIGURE 8. Precision.

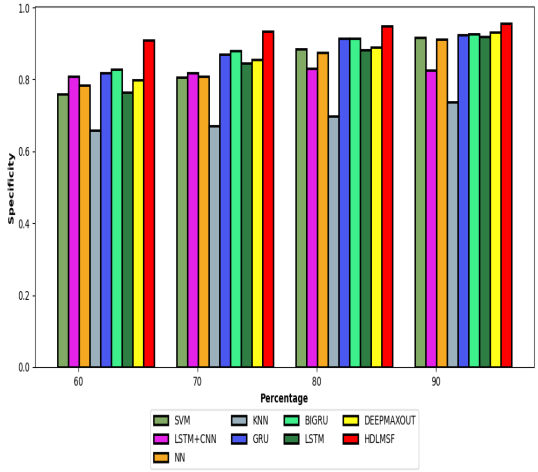

FIGURE 10. Sensitivity.

3) ASSESSMENT ON HDLMSF AND CONVENTIONAL STRATEGIES REGARDING NEGATIVE METRIC FOR JUDGMENT PREDICTION

In the context of a confusion matrix, negative evaluation metrics focus on the model's ability to correctly identify and handle the negative class (True Negatives, TN). These metrics are critical in domains where it is important to avoid False Positives, as incorrectly predicting negatives as positives can lead to serious consequences. Negative Prediction Value(NPV) measures the proportion of predicted negatives that are actual negatives. A low NPV means many false negatives, which can lead to overlooking important positive cases. False Negative Rate (FNR) measures the proportion of positive cases incorrectly classified as negative. While related to the positive class, FNR also complements negative evaluation metrics by showing how often positive cases are misclassified, which impacts negative predictions indirectly. A low FNR value ensured that the model is doing well. False Positive Rate(FPR) measures the proportion of negative cases incorrectly classified as positive. A low FPR means the model is performing well on negatives. For the training rate 90, the FNR of the HDLMSF methodology is 2.8553 which is very low when compare to other baseline models. Likewise, for all training rates, HDLMSF achieved fewer FNR and FPR error ratings. Figures 11 and 12 show the FPR and FNR values.

4) ASSESSMENT ON HDLMSF AND CONVENTIONAL STRATEGIES REGARDING OTHER METRICS FOR JUDGMENT PREDICTION

The other metric evaluations, such as F-measure, MCC, and NPV on HDLMSF and the traditional schemes for judgment prediction, is displayed in Figure 13, 14, 15. F_measure score balances Precision and Recall, making it critical when both False Positives and False Negatives are significant. Matthews Correlation Coefficient(MCC) is a balanced metric that evaluates the quality of predictions for both positive

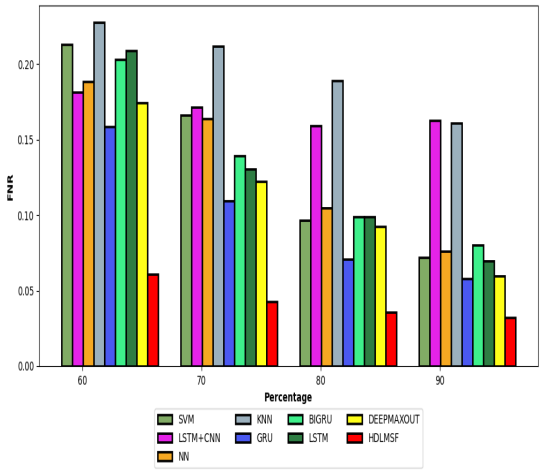

FIGURE 11. FNR.

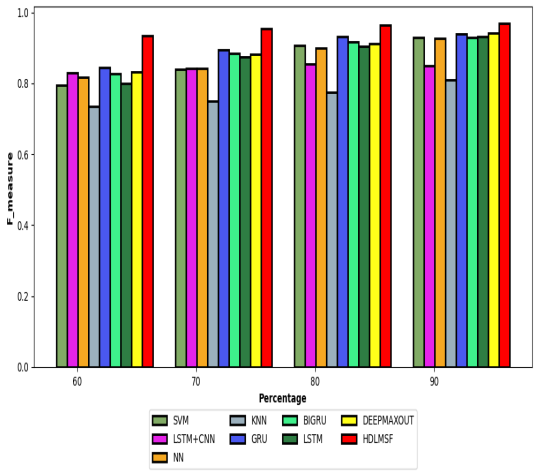

FIGURE 13. F Measure.

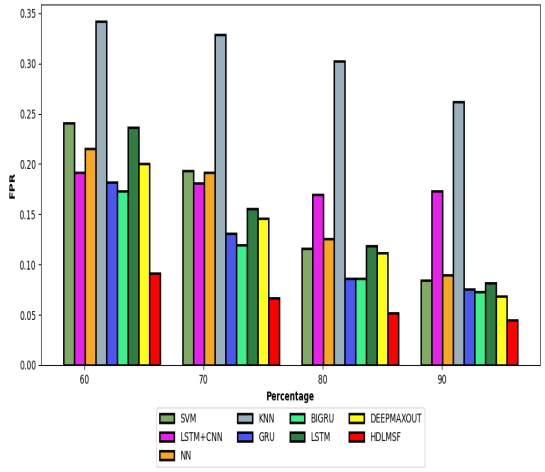

FIGURE 12. FPR.

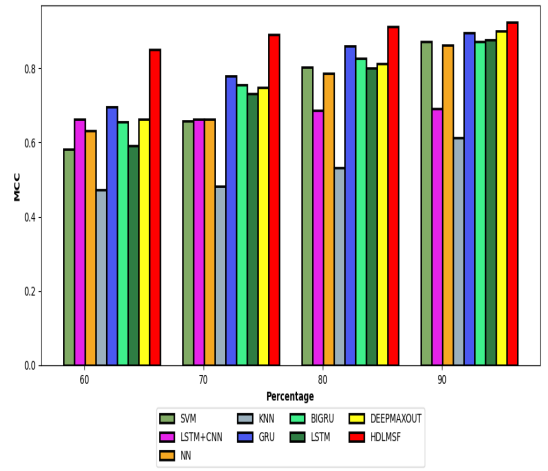

FIGURE 14. MCC.

and negative classes, even in imbalanced datasets. Negative Predictive Value(NPV) indicates the reliability of negative predictions, which is crucial in scenarios where False Negatives are dangerous. The F-measure of the HDLMSF approach is 92.1272 at 60% training percentage, MCC of the HDLMSF scheme is 93.2989, thereby offering better performance value when compared with the rest of the models.

5) IMPACT ON HDLMSF, MODEL WITH CONVENTIONAL ELMO, AND MODEL WITH CONVENTIONAL SCORE-LEVEL FUSION FOR JUDGMENT PREDICTION

The impact on HDLMSF, the model with conventional ELMo, and the model with conventional score-level fusion for judgment prediction is shown in Table 2. Here, the HDLMSF scheme with improved ELMo and score-level fusion has achieved superior values for the more accurate prediction of judgment. In particular, the NPV of the HDLMSF methodology is 0.9344, the model with conventional ELMo is 0.8750, and the model with conventional score level fusion is 0.8932, correspondingly. In addition, the HDLMSF acquired the FNR of 0.0476, model with conventional ELMo is 0.1071 and model with conventional score level fusion is 0.0893. Mainly, the HDLMSF yielded the accuracy = 0.9416, f-measure = 0.9496, FPR = 0.0733 and sensitivity = 0.9524. Altogether, the betterment of the HDLMSF with improved ELMo and improved score level fusion has produced enhanced outcomes. Following figure 16 visualizes the line plot of HDLMSF with traditional ELMo and traditional normalized score fusion. The above line plot interprets that HDLMSF with an ensemble word embedding feature, improved self-attention mechanism and modified score level fusion performs well in terms of accuracy, sensitivity, specificity, MCC, and F-measure when compared with traditional models. Also a comparison on transformer models such as BERT, Legal BERT and XLNet yields less accuracy on inputids and attention masks of lemmatized

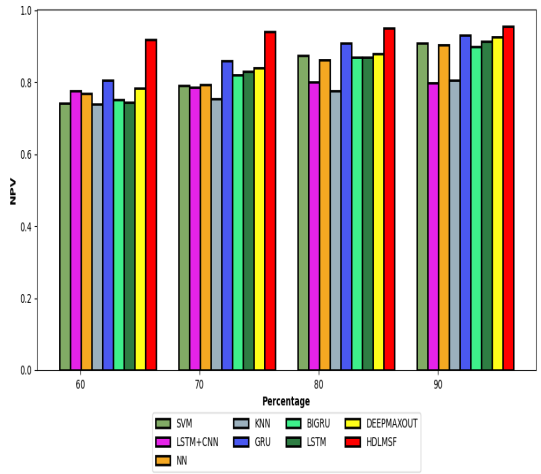

FIGURE 15. NPV.

TABLE 3. Ablation study on HDLMSF for judgment prediction.

Metrics	HDLMSF without ELMo + USE	HDLMSF without modified score-level fusion	HDLMSF with ELMo + USE and modified score fusion
Accuracy	0.8820	0.9002	0.9416
FNR	0.1071	0.0893	0.0476
Specificity	0.8673	0.8858	0.9267
F-measure	0.8900	0.9079	0.9496
FPR	0.1327	0.1142	0.0733
MCC	0.8198	0.8403	0.8801
Precision	0.8871	0.9052	0.9467
NPV	0.8750	0.8932	0.9344
Sensitivity	0.8929	0.9107	0.9524

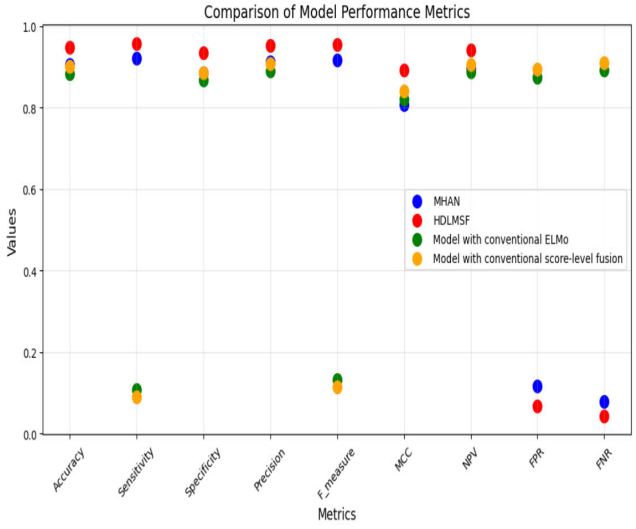

FIGURE 16. Comparison of model performance metrics.

documents 17. This shows that the reduction in accuracy of transformer models is due to the need for fine-tuning of hyperparameters to be done over billions of parameters present in the transformer models to improve the accuracy. Comparatively, Legal BERT shows a better score among transformer models as it is pretrained on legal words used in European English cases.

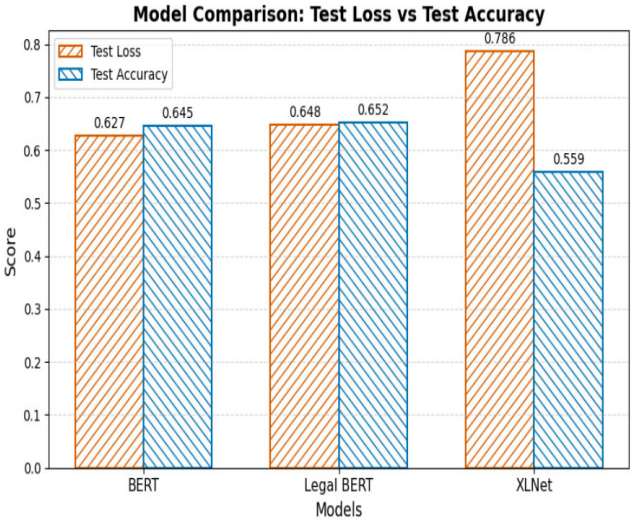

FIGURE 17. Comparison of BERT, LegalBERT and XLNet.

6) STATISTICAL ASSESSMENT ON HDLMSF AND CONVENTIONAL STRATEGIES WITH REGARD TO ACCURACY METRIC FOR JUDGMENT PREDICTION

Statistical evaluation is in need to validate the stability of the analysis. Accordingly, the statistical evaluation on HDLMSF is computed against the SVM, LSTM+CNN, NN, KNN, GRU, Bi-GRU, LSTM and Deep Maxout for judgment prediction is illustrated in Table 4 and shown in figure 18. Moreover, it is examined with regard to the accuracy measures under a distinctive type of statistical metrics. Further, the HDLMSF has recorded the highest accuracy metric ratings in almost all the statistical measures. Particularly, the accuracy of the HDLMSF is 0.9676 at the maximum statistical measure, while the conventional strategies gained minimal accuracy values, notably, SVM = 0.9230, LSTM+CNN = 0.8365, NN = 0.9185, KNN = 0.7923, GRU = 0.9347, Bi-GRU = 0.9229, LSTM = 0.9253 and Deep Maxout = 0.9366, correspondingly. Additionally, the HDLMSF accomplished the greatest accuracy value of 0.9471 under the median statistical metric, mean while the SVM, LSTM+CNN, NN, KNN, GRU, Bi-GRU, LSTM, and Deep Maxout scored the least accuracy values. In almost all statistical metrics, the HDLMSF model has exhibited its betterment by attaining greater accuracy. This has paved the way for accurate prediction of judgments.

In order to improve feature representation and model stability, future improvements to the current judgment prediction framework might investigate the merging of self-supervised learning and graph contrastive learning. When modeling the intricate relational structures seen in legal documents, such as precedent links, statute-to-fact linkages, and hierarchical

TABLE 4. Statistical study on HDLMSF and traditional methodologies regarding accuracy metric for judgment prediction.

Methods	Median	Minimum	Mean	Standard Deviation	Maximum
SVM	0.8586	0.7749	0.8538	0.0677	0.9230
LSTM+CNN	0.8285	0.8139	0.8269	0.0100	0.8365
NN	0.8552	0.7998	0.8572	0.0547	0.9185
KNN	0.7438	0.7168	0.7492	0.0331	0.7923
GRU	0.9022	0.8309	0.8925	0.0469	0.9347
Bi-GRU	0.8877	0.8096	0.8770	0.0503	0.9229
LSTM	0.8759	0.7790	0.8640	0.0628	0.9253
Deep Maxout	0.8836	0.8146	0.8796	0.0517	0.9366
HDLMSF	0.9471	0.9105	0.9431	0.0242	0.9676

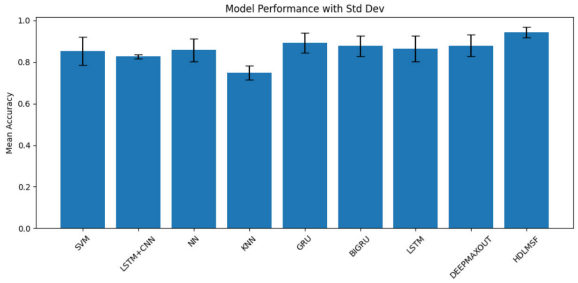

FIGURE 18. Statistical significance test.

case dependencies, graph contrastive learning [47] can be especially helpful. Beyond sequential text patterns, models are able to capture deeper semantic associations by learning invariant representations across augmented graph views. By pretraining models on vast amounts of unlabeled legal text, self-supervised learning can also overcome the problems of label sparsity and data imbalance. This enables models to learn contextual representations that generalize effectively even with a small number of labeled examples. These methods have demonstrated great promise in other high-stakes fields [48]. Scalability and optimization are two important factors that need to be considered, according to this work. Since the dataset used is small, a map-reduce framework suggested in [49] could be used to experiment with new models on judgment prediction. Also, fine-tuning the parameter using metaheuristic optimization [50] or computational optimization [51] would surely enhance the accuracy of the deep learning models for the prediction task.

E. CONCLUSION

Legal text analysis is rapidly emerging as a critical area of research, driven by the growing availability of digital legal content. The proposed HDLMSF model has been developed and evaluated using real-time Madras High Court criminal case data, demonstrating superior performance over several baseline classifiers, including SVM, KNN, BiGRU, DeepMaxout, GRU, LSTM, LSTM+CNN, neural networks, and transformer-based models such as BERT, Legal BERT, and XLNet. Because Bi-GRU models are lighter and more cost-effective than transformer-based models in terms of computational complexity and parameter utilization, this research work recommends their use. Furthermore, for small datasets, BiGRU performs better with less overfitting. The greater self-attention introduced to the Bi-GRU addresses the usual limitation of BiGRUs in capturing global context. On tiny datasets, transformer models may need extensive fine-tuning and may have trouble handling out-of-distribution data. In contrast to BiGRU, hierarchical BERT chunks the document, which could break the semantic flow in longer sequences. Additionally, the experimental findings on Legal BERT, BERT, and XLNet show that the proposed HDLMSF technique has a higher prediction accuracy than these models, as they needed to be fine-tuned more to give better accuracy.

This improved performance can be attributed to key innovations in the model design, including the use of ensemble word embeddings (ELMo+USE) with dimensionality reduction via an enhanced Principal Component Analysis, a modified self-attention mechanism, and score-level fusion within a hybrid classifier framework. The modified normalized fusion approach adopted effectively mitigates the influence of outliers, addressing a common limitation in existing methods. Despite these advancements, several challenges persist. These include the complexity of preprocessing legal case facts, which can vary significantly across jurisdictions the issue of similar cases leading to divergent judgments, and the need for extensive hyperparameter tuning in transformer models.

Future work will focus on refining the embedding techniques and optimizing model performance through the integration of Large Language Models. In addition, a critical area of concern is the potential for legal AI systems to inherit and perpetuate systemic biases present in historical data, particularly those related to gender, socioeconomic status, or race. Without careful auditing and fairness interventions, such systems risk reinforcing discriminatory outcomes under the guise of objectivity. Therefore, future research will also prioritize the development of explainability frameworks and bias mitigation strategies to ensure that predictive outcomes are not only accurate but also ethically and legally sound.

REFERENCES

[1] Y. Lyu, Z. Wang, Z. Ren, P. Ren, Z. Chen, X. Liu, Y. Li, H. Li, and H. Song, "Improving legal judgment prediction through reinforced criminal element extraction," *Inf. Process. Manage.*, vol. 59, no. 1, Jan. 2022, Art. no. 102780, doi: 10.1016/j.ipm.2021.102780.

[2] P. Madambakam, S. Rajmohan, H. Sharma, and T. Anka Chandrahas Purushotham Gupta, "SLJP: Semantic extraction based legal judgment prediction," 2023, *arXiv:2312.07979*.

[3] Y.-X. Hong and C.-H. Chang, "Improving colloquial case legal judgment prediction via abstractive text summarization," *Comput. Law Secur. Rev.*, vol. 51, Nov. 2023, Art. no. 105863, doi: 10.1016/j.clsr.2023.105863.

[4] X. Guo, L. Zhang, and Z. Tian, "Judgment prediction based on tensor decomposition with optimized neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 8, pp. 11116–11127, Aug. 2024, doi: 10.1109/TNNLS.2023.3248275.

[5] P. Li, A. Luo, J. Liu, Y. Wang, J. Zhu, Y. Deng, and J. Zhang, "Bidirectional gated recurrent unit neural network for Chinese address element segmentation," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 11, p. 635, Oct. 2020, doi: 10.3390/ijgi9110635.

[6] K. Javed and J. Li, "Artificial intelligence in judicial adjudication: Semantic biasness classification and identification in legal judgement (SBCILJ)," *Heliyon*, vol. 10, no. 9, May 2024, Art. no. e30184, doi: 10.1016/j.heliyon.2024.e30184.

[7] J. K. Nuamah and Y. Seong, "A machine learning approach to predict human judgments in compensatory and noncompensatory judgment tasks," *IEEE Trans. Human-Mach. Syst.*, vol. 49, no. 4, pp. 326–336, Aug. 2019, doi: 10.1109/THMS.2019.2892436.

[8] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, "How does NLP benefit legal system: A summary of legal artificial intelligence," 2020, *arXiv:2004.12158*.

[9] C. Xiao, H. Zhong, Z. Guo, C. Tu, Z. Liu, M. Sun, Y. Feng, X. Han, Z. Hu, H. Wang, and J. Xu, "CAIL2018: A large-scale legal dataset for judgment prediction," 2018, *arXiv:1807.02478*.

[10] T. Dolci, F. Azzalini, and M. Tanelli, "Improving gender-related fairness in sentence encoders: A semantics-based approach," *Data Sci. Eng.*, vol. 8, no. 2, pp. 177–195, Jun. 2023, doi: 10.1007/s41019-023-00211-0.

[11] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos, "Predicting judicial decisions of the European court of human rights: A natural language processing perspective," *PeerJ Comput. Sci.*, vol. 2, p. e93, Oct. 2016, doi: 10.7717/peerj-cs.93.

[12] M. Medvedeva, M. Vols, and M. Wieling, "Using machine learning to predict decisions of the European court of human rights," *Artif. Intell. Law*, vol. 28, no. 2, pp. 237–266, Jun. 2020, doi: 10.1007/s10506-019-09255-y.

[13] O.-M. Sulea, M. Zampieri, M. Vela, and J. van Genabith, "Predicting the law area and decisions of French supreme court cases," 2017, *arXiv:1708.01681*.

[14] J. Niklaus, I. Chalkidis, and M. Stürmer, "Swiss–Judgment-prediction: A multilingual legal judgment prediction benchmark," 2021, *arXiv:2110.00806*.

[15] D. Wei and L. Lin, "An external knowledge enhanced multi-label charge prediction approach with label number learning," 2019, *arXiv:1907.02205*.

[16] N. Xu, P. Wang, L. Chen, L. Pan, X. Wang, and J. Zhao, "Distinguish confusing law articles for legal judgment prediction," 2020, *arXiv:2004.02557*.

[17] *High Court Judges Get Just 5-6 Minutes to Decide Cases, Says Study*. Accessed: Jan. 17, 2025. [Online]. Available: https://timesofindia.indiatimes.com/india/high-court-judges-get-just-5-6-minutes-to-decide-cases-says-study/articleshow/51722614.cms

[18] *Decoding Delay: Analysis of Court Data*. Accessed: Feb. 12, 2025. [Online]. Available: https://www.dakshindia.org/Daksh_Justice_in_India/19_chapter_01.xhtml

[19] H. Elfaik and E. H. Nfaoui, "Deep bidirectional LSTM network learning-based sentiment analysis for Arabic text," *J. Intell. Syst.*, vol. 30, no. 1, pp. 395–412, Dec. 2020, doi: 10.1515/jisys-2020-0021.

[20] Y. Dong, Y. Gong, X. Bo, and Z. Tan, "Early quality prediction of complex double-walled hollow turbine blades based on improved whale optimization algorithm," *J. Comput. Inf. Sci. Eng.*, vol. 25, no. 1, p. 25, Jan. 2025.

[21] G. Sukanya and J. Priyadarshini, "Analysis on word embedding and classifier models in legal analytics," *AIP Conf. Proc.*, vol. 3040, no. 1, 2024, Art. no. 140001, doi: 10.1063/5.0181820.

[22] S. Abbara, M. Hafez, A. Kazzaz, A. Alhothali, and A. Alsolami, "ALJP: An Arabic legal judgment prediction in personal status cases using machine learning models," 2023, *arXiv:2309.00238*.

[23] G. Sukanya and J. Priyadarshini, "Modified hierarchical-attention network model for legal judgment predictions," *Data Knowl. Eng.*, vol. 147, Sep. 2023, Art. no. 102203, doi: 10.1016/j.datak.2023.102203.

[24] Z. Chen, Y. Bao, and T. Zhu, "An empirical study on IPO model construction of undergraduate education quality evaluation in China from the statistical pattern recognition approach in NLP," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 21, no. 6, pp. 1–15, Nov. 2022, doi: 10.1145/3543851.

[25] V. G. F. Bertalan and E. E. S. Ruiz, "Using attention methods to predict judicial outcomes," *Artif. Intell. Law*, vol. 32, no. 1, pp. 87–115, Mar. 2024, doi: 10.1007/s10506-022-09342-7.

[26] S. Kumar Nigam, A. Deroy, S. Maity, and A. Bhattacharya, "Rethinking legal judgement prediction in a realistic scenario in the era of large language models," 2024, *arXiv:2410.10542*.

[27] J. Kong, J. Wang, and X. Zhang, "Hierarchical BERT with an adaptive fine-tuning strategy for document classification," *Knowl.-Based Syst.*, vol. 238, Feb. 2022, Art. no. 107872, doi: 10.1016/j.knosys.2021.107872.

[28] P. T. Huu, N. T. An, N. N. Trung, H. N. Thien, N. S. Duc, and N. T. Ty, "Judicial decision prediction using an integrated attention based bidirectional long-short term memory and dilated skip residual convolution neural network," *Vis. Comput.*, vol. 41, no. 6, pp. 4199–4220, Apr. 2025.

[29] J. Chen, X. Zhang, X. Zhou, Y. Han, and Q. Zhou, "An approach based on cross-attention mechanism and label-enhancement algorithm for legal judgment prediction," *Mathematics*, vol. 11, no. 9, p. 2032, Apr. 2023, doi: 10.3390/math11092032.

[30] S. Shen, G. Qi, Z. Li, S. Bi, and L. Wang, "Hierarchical Chinese legal event extraction via pedal attention mechanism," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 100–113, doi: 10.18653/v1/2020.coling-main.9.

[31] Q. Bao, H. Zan, P. Gong, J. Chen, and Y. Xiao, "Charge prediction with legal attention," in *Natural Language Processing and Chinese Computing*. Cham, Switzerland: Springer, 2019, doi: 10.1007/978-3-030-32233-5_35.

[32] A. Shelar and M. Moharir, "Judgment prediction from legal documents using Texas wolf optimization based deep BiLSTM model," *Intell. Decis. Technol.*, vol. 18, no. 2, pp. 1557–1576, Jun. 2024, doi: 10.3233/idt-230566.

[33] Y. Dong, X. Li, J. Shi, Y. Dong, and C. Chen, "Graph contrastive learning networks with augmentation for legal judgment prediction," *Artif. Intell. Law*, vol. 10, no. 5, Jun. 2024, doi: 10.1007/s10506-024-09407-9.

[34] N. Xu, P. Wang, J. Zhao, F. Sun, L. Lan, J. Tao, L. Pan, and X. Guan, "Distinguish confusion in legal judgment prediction via revised relation knowledge," 2024, *arXiv:2408.09422*.

[35] X. Guo, H. Zhang, L. Ye, S. Li, and G. Zhang, "TenRR: An approach based on innovative tensor decomposition and optimized ridge regression for judgment prediction of legal cases," *IEEE Access*, vol. 8, pp. 167914–167929, 2020, doi: 10.1109/ACCESS.2020.2999522.

[36] H. Zhang, Z. Dou, Y. Zhu, and J.-R. Wen, "Contrastive learning for legal judgment prediction," *ACM Trans. Inf. Syst.*, vol. 41, no. 4, pp. 1–25, Oct. 2023.

[37] Y. Zhang, W. Huang, Y. Feng, C. Li, Z. Fei, J. Ge, B. Luo, and V. Ng, "LJPCheck: Functional tests for legal judgment prediction," in *Proc. Findings Assoc. Comput. Linguistics ACL*, 2024, pp. 5878–5894, doi: 10.18653/v1/2024.findings-acl.350.

[38] A. Haidar, T. Ahajjam, I. Zeroual, and Y. Farhaoui, "Application of machine learning algorithms for predicting outcomes of accident cases in Moroccan courts," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 26, no. 2, p. 1103, May 2022, doi: 10.11591/ijeecs.v26.i2.pp1103-1108.

[39] D. Alghazzawi, O. Bamasag, A. Albeshri, I. Sana, H. Ullah, and M. Z. Asghar, "Efficient prediction of court judgments using an LSTM+CNN neural network model with an optimal feature set," *Mathematics*, vol. 10, no. 5, p. 683, Feb. 2022, doi: 10.3390/math10050683.

[40] X. Shang, "A computational intelligence model for legal prediction and decision support," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–8, Jun. 2022, doi: 10.1155/2022/5795189.

[41] J. Seo, S. Lee, L. Liu, and W. Choi, "TA-SBERT: Token attention sentence-BERT for improving sentence representation," *IEEE Access*, vol. 10, pp. 39119–39128, 2022, doi: 10.1109/ACCESS.2022.3164769.

[42] *Guide to Learn ELMo for Extracting Features from Text*. Accessed: Feb. 14, 2025. [Online]. Available: https://www.analyticsvidhya.com/blog/2019/03/learn-to-use-elmo-to-extract-features-from-text/

[43] P. Hall, H.-G. Müller, and J.-L. Wang, "Properties of principal component methods for functional and longitudinal data analysis," *Ann. Statist.*, vol. 34, no. 3, pp. 1493–1517, Jun. 2006, doi: 10.1214/009053606000000272.

[44] M. S. Akhtar, T. Garg, and A. Ekbal, "Multi-task learning for aspect term extraction and aspect sentiment classification," *Neurocomputing*, vol. 398, pp. 247–256, Jul. 2020, doi: 10.1016/j.neucom.2020.02.093.

[45] J. Peta and S. Koppu, "An IoT-based framework and ensemble optimized deep maxout network model for breast cancer classification," *Electronics*, vol. 11, no. 24, p. 4137, Dec. 2022, doi: 10.3390/electronics11244137.

[46] D. Alghazzawi, O. Bamasag, A. Albeshri, I. Sana, H. Ullah, and M. Z. Asghar, "Efficient prediction of court judgments using an LSTM+ CNN neural network model with an optimal feature set," *Mathematics*, vol. 10, no. 683, p. 683, 2022, doi: doi: 10.3390/math10050683.

[47] X. Yang, G. Yang, and J. Chu, "GraphCL-DTA: A graph contrastive learning with molecular semantics for drug-target binding affinity prediction," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 8, pp. 4544–4552, Aug. 2024, doi: 10.1109/JBHI.2024.3350666. [Online]. Available: https://api.semanticscholar.org/CorpusID:259950801

[48] X. Yang, G. Yang, and J. Chu, "Self-supervised learning for label sparsity in computational drug repositioning," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 20, no. 5, pp. 3245–3256, Sep. 2023, doi: 10.1109/TCBB.2023.3254163. [Online]. Available: https://api.semanticscholar.org/CorpusID:249240152

[49] K. M. Matrouk, J. E. Nalavade, S. Alhasen, M. Chavan, and N. Verma, "MapReduce framework based sequential association rule mining with deep learning enabled classification in retail scenario," *Cybern. Syst.*, vol. 56, no. 2, pp. 147–169, Feb. 2025, doi: 10.1080/01969722.2023.2166256. [Online]. Available: https://api.semanticscholar.org/CorpusID:257251151
[50] G. Sukanya and J. Priyadarshini, "Fine tuned hybrid deep learning model for effective judgment prediction," *Comput. Model. Eng. Sci.*, vol. 142, no. 3, pp. 2925–2958, 2025, doi: 10.32604/cmes.2025.060030. [Online]. Available: https://api.semanticscholar.org/CorpusID:276766750
[51] D. T. Dang and N. T. Nguyen, "New evolutionary algorithms for determining consensus of ordered partition collectives," *Cybern. Syst.*, pp. 1–24, Jan. 2024, doi: 10.1080/01969722.2023.2296247.
[52] L. Laws. (2023). *A Judge in a High Court Spends Just Five Minutes Hearing a Case, Reveals Study*. LatestLaws.com. [Online]. Available: https://www.latestlaws.com/latest-news/a-judge-in-a-high-court-spends-just-five-minutes-hearing-a-case-reveals-study/?utm_source=chatgpt.com
[53] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

G. SUKANYA received the B.E. degree in computer science engineering from Madras University, Chennai, India, in 2004, and the M.E. degree in computer science and engineering from Anna University, India, in 2008. She is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, VIT University, Chennai. She was an Assistant Professor at the Alpha College of Engineering for around six years. Her favourite subjects are artificial intelligence, cryptography, and data science, where she has hands-on experience. She has published research articles related to attention models on legal judgment prediction, encryption algorithms, and image recognition. Her research interests include artificial intelligence, machine learning, natural language processing, and legal analytics. She spends most of her time in legal research by visiting and understanding the problems faced in that area in Indian court cases.

J. PRIYADARSHINI received the B.E. and M.Tech. degrees in computer science and engineering from Anna University, Chennai, Tamil Nadu, India, in 2005 and 2008, respectively, and the Ph.D. degree in computer science from Anna University, MIT Campus, Chennai, in 2014. She is currently a Professor with the School of Computer Science and Engineering, VIT University, Chennai. She has also published many papers in various national and international conferences and journals. Her research interests include artificial intelligence, image processing, and natural language processing.