

*Iqbal et al. LGURJCSIT 2025*

*LGU Research Journal of Computer Science & IT*

*ISSN: 2521-0122 (Online)*
*ISSN: 2519-7991 (Print)*

*doi: 10.54692/lgurjcsit.2025.092685*

*Vol (9): Issue (2), July – December 2025*

# Predictive Modeling of Court Decisions using Transformer-Based NLP and Deep Learning Techniques

Ayesha Iqbal[1], Hina Sattar[1], Aqsa Mahmood[1], Benish Nadeem[1], Muhammad Munwar Iqbal[2*]
[1]Department of Computer Science & IT, Government Sadiq College Women University, Bahawalpur, 63100, Punjab, Pakistan
[2]Department of Computer Science, University of Engineering and Technology Taxila, Taxila, Pakistan.

Email: munwariq@gmail.com

**ABSTRACT:**

***The high growth in digital legal documents has brought about immense tasks in information retrieving, categorizing, and decision-making processes about the law. Manual legal analysis is still time-consuming, inconsistent and heavily resource-dependent. Thus, there is a requirement for a transition to automated LegalTech. The current study examines how legislative acts in the European Union (EU) can be predicted based on legal decisions made using Natural Language Processing (NLP) and hybrid deep learning. Based on the use of the EurLex corpus with more than 142,000 legislative texts, this study focuses on classifying three major classes: Directives, Regulations, and Decisions. We employ a multi-stage pipeline in order to deal with not only the super-ordinary linguistic diversity of the legal texts but also the imbalance between the masses and categories. We propose and test composite deep learning models that join the Convolutional Neural Networks (CNNs), the Long Short-term memory (LSTM) and the Gated Recurrent Unit (GRU). These have been improved by adding attention mechanisms in order to cover both long-distance and local phrase structures. Its semantical interpretation is also reinforced in the model through FastText embedding that captures the subword law vocabulary. The findings of our experiment suggest that hybrid models are more accurate, recall, and F1-score than the traditional baselines of machine learning, especially in terms of multi-class classification of the legislative. It is enhanced by the incorporation of structured metadata, including Act Type and Subject Matter, that adds meaningful value to feature representation and generates a higher level of interpretability. This study focused on legislative data, while ignoring judicial data. It contributes to the use of AI in legal analytics, making a case for a scalable and explainable document classification approach that transcends linguistic and jurisdictional boundaries.***

**KEYWORDS:** Legal decisions, CNN-LSTM architecture, GRU-LSTM architecture, FastText features, EurLex corpus, Categorizing of summary legislative documents.

## 1. INTRODUCTION

A digital era has also triggered the volume and complexity of legal acts to expand exponentially, which has been quite challenging for legal experts, professionals, and even policymakers. Once it opens, traditional and manual legal research is slow, expensive, prone to error, and subject to subjectivity. To that end, Legacy has risen to address this market need by utilising technology to automate legal processes, manage documents, and provide insight into decisions. The most striking feature of such a transformation is the integration of Artificial Intelligence (AI) and Natural Language Processing (NLP) into the

great body of legal texts, known as retiling informative trends and prognostic analytics in multilateral legal textual prose [1]. Legal text classification and decision prediction entail predicting whether a legal text, like a court case result, a statutory decision or an act of parliament, has or will have a particular outcome/category based on the linguistic and contextual content of the text [2]. This is never an easy task, owing to the lengthy, technical, and situational language of the law. Moreover, Jurisdictional and legal application legal documents come in a variety of formats, lengths and structures.

Although it has been demonstrated that hybrid deep learning architectures consisting of Convolutional Neural Networks (CNNs) to extract local features (e.g., legal phrasing) and Recurrent Neural Networks (RNNs) to analyze them sequentially (e.g., Long Short-Term Memory, LSTM networks) address important research gaps [4, 5, 6], still, this is far from everything. Existing literature is mainly concerned with judicial sources (e.g. court decisions in Indonesia or the U.S. [5, 6]) and mostly faces difficulties that relate to issues of data imbalance, the absence of both multilingual and domain specific embeddings and the generalizing from one legal system to another [1]. It illustrates the necessity to investigate other multilingual datasets in other fields than the judicial field to facilitate the area of legal AI.

This paper fills this gap as it does not rely on judicial writings but concentrates on categorizing the legislative texts of the European Union. The authors' first aim is to construct and test a strong predictive model to classify legal acts in the vast EurLex contents [7]. The system includes more than 142,000 legally binding EU acts (such as Regulations, Directives, Decisions) which have been adopted since 1952. To achieve the above goal, we suggest a new hybrid deep-learning architecture that combines CNN and LSTM networks. This model is enhanced with FastText embeddings to better capture subword-level semantics and legal jargon, an attention mechanism to improve interpretability by focusing on relevant text segments, and structured metadata to provide additional contextual information. Experiments we conducted show that the suggested model is able to compete on the EurLex data at the state of the art, be it compared to the standard baseline models. The study makes the following contributions:

- We develop a hybrid CNN-LSTM model augmented with FastText embeddings and attention layers for superior semantic representation and classification of legislative texts.
- We innovatively integrate structured metadata with raw text to provide rich contextual cues, leading to more accurate and interpretable predictions.
- We generalized legal AI research to the legislative domain by employing the extensive, multilingual EurLex corpus, pioneering large-scale predictive analytics for EU policy and lawmaking.

The following is the structure of this paper. The second Section II presents a close literature review on the work in terms of legal predicting decision, deep learning networks and the texts of data in NLP. Section III explains the proposed methodology, which consists of a data collection and preprocessing phase, the design of the hybrid CNN-LSTM model, the usage of FastText embeddings and attention mechanisms, and the use of structured metadata of the EurLex corpus. In Section IV, the experimental methodology, evaluation criteria and findings are elaborated, and then the performance of the model against various types of law and legal attributes will be analyzed in detail. Lastly, Section V concludes the paper summarizing the most important findings and explaining the practical implications and the possible direction of further research in the field of legal AI and document classification.

## 2. LITERATURE REVIEW

Legal Decision Prediction (LDP), also known as Legal Natural Language Processing (NLP), has become an important line of research within Legal NLP with the primary goal being to predict the outcome of a legal case through the usage of textual data that is derived from judgments, statutes, and regulatory documents [9]. On the one hand, the trend of the digitization of legal texts has resulted in raising more interest in developing computational tools to assist legal professionals in their everyday activities, like precedents retrieval, sentence proposals, and prediction of legal outcomes [24]. On the other hand, in the early days of LDP, various classical machine learning algorithms were mainly used, including Support Vector Machines (SVM), Logistic Regression, and Decision Trees [26]. These approaches were also very dependent on hand-coded features, ngrams, keyword extraction techniques and domain-specific legal indicators, which boasted low levels of generalizability across jurisdictions and failed to capture the linguistic complexity of

legal documents [10]. With the introduction of deep learning, this domain has changed as it is now possible to implement automatic feature extraction and better semantic representation. Convolutional Neural Networks (CNNs) were found to be practical in detecting the patterns that come with syntax in the smaller scale, the recurrent neural networks (RNNs), especially the long short-term memory (LSTM) and the BiLSTM proved to be effective in capturing the tie of sequential and hierarchical values that have the manifestation of the trend of the persuasiveness of the law [25]. More recently, architectures using transformers as BERT and domain-adapted versions thereof (e.g. LegalBERT, CaseLaw-BERT) have pushed the state-of-the-art in legal text understanding [2]. The models take advantage of self-attention capabilities in order to represent long-range dependencies and have been adapted on legal corpora to better suit the domain-specific vocabulary and the context of discussion.

The combined use of CNNs, LSTMs and attention-based models in architecture has been shown to be of much potential as hybrid deep learning models. Remarkably, Ibrahim et al. developed a CNN-BiLSTM-attention model to predict Indonesian court rulings and indicated an improvement in predictive accuracy and interpretability through visualization of attention physiology [11]. Sukanya and Priyadarshini reported a CNN-transformer combination model that reached state-of-the-art performance on benchmark criminal case outcome prediction datasets used a combination of rule-based reasoning and deep learning and reported F1-scores of more fthan 90% on a range of multilabel legal classification tasks. Overall, these studies highlight the act that hybrid models, due to the ability to represent syntactic sophistication and cultural ties, have superior legal NLP functionality over singular neural structures [12]. Annotated legal datasets are extremely important in the successful performance of legal decision prediction (LDP) models. There are a few benchmark corpora that have developed over the years, each brings in its own advantages, but shows a significant lack in one respect [27]. The EurLex corpus of more than 142,000 legislation items (issued by European Union bodies and organizations) is a notable exception due to its rich multi-academy and multilabel labeling as well as the inclusion of circa 23 structured fields of metadata in such categories as the subject matter and legal act type [13]. This qualifies it especially to be used in cross-jurisdictional legal text classification. Conversely, the JUSTICE dataset centres on legal rulings of the U.S Supreme Court, and allows the outcome prediction tasks. But its coverage is limited to one jurisdiction and not the covering of texts of legislation [14]. The LexGLUE standard combines a number of English-language legal datasets to facilitate judgment prediction, classification and summarization tasks. Although it allows assessing the models in a standard way, it is monolingual and judicial-centred and can thus be used in the wider setting of legislation [8]. Other corpora, like Court view generation and Chinese charge prediction datasets, are suitably designed when it comes to legal opinion generation in Chinese and are usually language-specific and do not house much metadata. There remain

***Table 1: Summary of Related Work in Legal Document Classification***

| References | Dataset Used | Methodology | Key Findings | Limitations |
|---|---|---|---|---|
| [22] | European Court of Human Rights (ECHR) | CNN | Effective for extracting legal facts from case text. | Focused only on judicial opinions, not legislative texts. Limited to a single court jurisdiction. |
| [23] | Indonesian Supreme Court Rulings | CNN-BiLSTM with Attention | Achieved high accuracy and improved model interpretability for verdict prediction. | Focused on a specific national judiciary. Did not address class imbalance or integrate metadata. |
| [6] | Indian Criminal Case Reports | CNN-Transformer | Demonstrated state-of-the-art performance for outcome prediction in criminal cases. | Limited to criminal law and a single jurisdiction. Model does not leverage structured metadata. |
| [8] | EUR-Lex (Subset) | Legal-BERT | Showed the effectiveness of pre-trained transformers on legal texts for multi-label tasks. | Did not explore hybrid architectures (CNN+RNN). Computational expensive. Limited analysis on class imbalance |
| [9] | Multi-Legal (Multi-Jurisdictional) | Legal-Llama 2 (Fine-tuned LLM) | Achieved new SOTA on multiple legal NLP benchmarks, showing strong zero-shot capabilities. | Extremely high computational cost for training and inference. Requires massive resources, limiting accessibility and practicality. |

various issues despite the advancements made with these datasets. The current available resources only feature a small amount of multilingual and multi-jurisdictional, and hence lower cross-legal system generalization [17]. Also, it is not enough to utilize structured metadata, so that many models overlook useful contextual information that may be found in legal documents. The presence of imbalanced classes can be cited as one of its primary issues, especially in a multilabel classification scenario when some law categories are highly underrepresented [18]. Also, the primary importance is paid to the genres of judicial documents, that is, there is a relatively insufficient attention to legislative corpora that are equally significant in legal analytics. Achieving the aforementioned limitations is crucial in attaining solid and generalizable legal AI systems [15][19]. In comparison, whereas much effort has been focused on textual classification of the law, current research is generally fixed on judicial data in certain jurisdictions without regard to legislative texts and related imbalances in data. The related work has been summarized in Table 1 of the study, which comprises the key findings of the dataset methodology and limitations.

### *2.1. Research Gap*

Although significant setbacks have been made in the sphere of legal decision prediction (LDP), there is still a complex of important issues that have not yet been solved. A significant deficit is that existing models lack adequate multilingual and cross-jurisdictional flexibility in that they have often been modeled to utilize monolingual data as well as trained towards a specific legal framework. This significantly reduces the effectiveness of the multi-jurisdictional processes, such as the European Union, whose legal documentations are translated into various languages and regulated by various legal traditions. In addition, the use of legal metadata that includes document type, subject, and legal reference has not reached full potential, even though these provide contextual information of documents that can enhance the quality of a prediction by making it more accurate or interpretable. Another issue is related to the inexplicability of deep learning models, as with such an approach, legal professionals cannot trust or comprehend the justification of model results. Such unexplainability is a barrier to the successful implementation of AI systems in areas that are sensitive to the law. Lastly, the lack of a variety of representative data in terms of both legislative and judicial texts that cuts across multilingual contexts and jurisdictions hinders the prospect of achieving more comprehensive and generalizable frameworks of legal AI. This paper shall address some of these gaps by doing the following:

- This was done by designing a hybrid deep learning approach that integrated the CNN, LSTM, with attention mechanisms to enhance the effectiveness and explainability of the results when determining the classification of legal texts.
- Using FastText embeddings and the legal metadata structure of the EurLex corpus to enhance multilingual representation, semantics knowledge, and contextual awareness.
- The concentration on texts of legislation instead of judicial opinions, thus expanding the applicability of legal AI and the ability to classify broadly any document with legal relevance of many types.

## 3. PROPOSED METHODOLOGY

A systematic approach to methodology is applied in this paper to develop and test hybrid deep learning models that perform multi-label classification of documents from the EurLex corpus. The steps used in the workflow, depicted in Figure 1, include five significant steps: data collection and preprocessing, text representation by means of FastText embeddings, feature extraction, development of models (both the conventional machine learning and hybrid deep learning ones), and model performance evaluation [18]. All stages are selected to overcome the domain-specific issues in legal document classification, like domain-specific language, class imbalance, and semantical complexity. The methodology eludes the need to harness only linguistic or structural characteristics of legal materials to enhance classification precision, as well as interpretability.

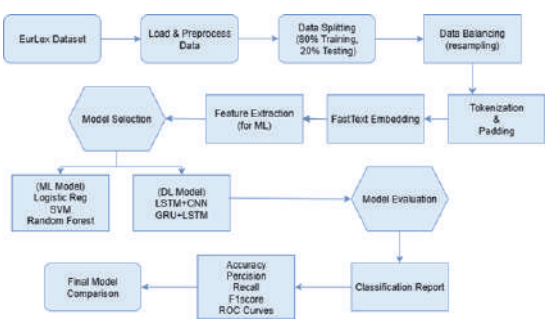



***Figure 1: Proposed Methodology Framework***

### *3.1. Data Collection and Preprocessing*

The EurLex dataset, which contains more than

142,000 legislative documents published by European Union institutions, has been downloaded via Kaggle. To fit the given research, three legislative categories have been analyzed, namely Directive, Regulation and Decision. The following preprocessing was carried out to ensure the quality of the input data:

• Language Filtering: It only kept the documents that are in the English language to ensure consistency in the training and evaluation of the model.

• Text Cleaning: This included lowercasing, cleaning up of punctuation marks, numbers and extraneous white space.

• Target Classes: The documents were grouped into three functional clusters i.e.: Directive, Regulation and Decision.

• Train-Test Split: In stratified sampling invoked on both training and testing sets the data was split into 80% training and 20% testing to maintain the class distribution found in the original subsets

• Handling Class Imbalance: The severe initial disparity (e.g., Regulations: 102,304 vs. Directives: 4,070) was saturated through random oversampling, only on the training set, with the RandomOverSampler tool of the imbalanced-learn library. This produced an ideal seed training of 99547 examples per class. The validation and test sets remained in their initial state of an imbalanced nature so that an adequate and objective repository of how models were to perform on actual real-world data is obtained.

### *3.2. Text Representation Using FastText embedding*

Pre-trained FastText word embeddings (wiki-news-300d-1M.vec) were employed in order to represent the semantic content of legal documents with a view to conveying such facts. In contrast to conventional embeddings like Word2Vec or GloVe, FastText is equipped to obtain subword-level features since the former uses the model of the word as a multi-set of character n-grams. It is particularly useful with law texts, as compound words and even jargon are very common. An embedding matrix that contains the 5,000 most common tokens in the corpus over a 300-dimensional space has been built. The implementation of the above embeddings has been as part of the deep learning model with the trainable=False option to ensure the semantic correctness of the already trained vectors remains the same during the learning exercise.

#### *A. Feature Preparation and Tokenization*

Textual data became prepared to be modeled as follows:

• Tokenization: The Karas Tokenizer was trained to create indexed sequences out of the words.

• Padding: Such padding of each sequence was completed to 100 tokens so that one can maintain consistency at the input.

• Label Encoding: Label Encoder encoding where the Label classes (Directive, Regulation, Decision) get converted to a one-hot encoded vector.

This preprocessing pipeline made sure that the input features were as well as the labels were appropriate for multi-classification.

#### *B. Feature Extraction*

There were two different ways of achieving sense out of the textual data:

• Sentence-Level Feature Extraction in machine learning models, sentence embeddings were composited as the mean of the FastText representation of individual words in each document, which gave fixed-length representations.

• With deep learning models, we simply indexed a series of tokens with their FastText embedding matrix. These interesting embeddings stored semantic and morphological context information, which was all-important to understanding legal language. These dual-level feature extractions helped the study to utilize the global document semantics as well as the fine-grained linguistic structures.

### *3.3. Machine Learning Model*

The section discusses how the models utilized in this study are structured. There were two different types of implemented models: traditional machine learning and hybrid deep learning. The deep learning models would then be built on top of the machine learning models as a way of ensuring that complex semantic and contextual dependencies would be included in the legal texts. In order to set a performance baseline, we applied some of the standard machine learning classifiers. These models have been developed on the basis of the sentence dimensions formed in the EurLex corpus using FastText. The algorithms that were employed are as follows:

• Logistic Regression: Probabilistic linear Model and deployed in classification. Logistic Regression refers to a probability: P (y | X), to the sigmoid or SoftMax probability. When one

considers a classification with more than 2 classes, then the formula given in Equation 1 is use by SoftMax.

$$p(y = j|X) = \frac{e_j^{\theta_j^T X}}{\sum_{k=1}^{K} e^{\theta_k^T X}} \; for \; j = 1, \ldots, K \quad \textbf{(1)}$$

• **Support Vector Machine:** It is a Probabilistic /Discriminative type of classifier that finds out the boundary hyperplane among classes using maximization of the distance between classes. Such a decision follows the rule given in Equation 2.

$$f(x)=(\omega^T x + b) \quad \textbf{(2)}$$

• **Random Forest:** Random forest is an ensemble algorithm using which multiple decision trees are constructed, and the mode of the predictions of the trees is delivered as output. The prediction ŷ is given by the formula in Equation 3.

$$\hat{y} = \text{majority_}vote(T_1\,(x), T_2\,(x), \ldots, T_n\,(x)) \quad \textbf{(3)}$$

All of the models have been trained on the sentence-level FastText embedding as the features. Their performance numbers will be used as comparative benchmarks to measure the performance of the presented hybrid deep learning models [20]. The traditional models are not computationally intense and are typically unable to capture the sequential relationship and complexity of context typical in complex legal language.

### *3.4. Hybrid Deep Learning model*

Two hybrid deep learning structures have been designed in order to learn both sequence dependence and local patterns of linguistics in legal texts. Each of these models uses Long Short-Term Memory (LSTM) layers either in conjunction with Convolutional Neural Networks (CNN) or Gated Recurrent Units (GRU), the bases of which enable the formation of more advanced granular semantics and improved classification results.

#### *A. Model 1: LSTM AND CNN*

It is a hybrid model in which an LSTM layer is incorporated with the CNN. The LSTM component takes in long-range dependency and sequence context in legal documents, which is necessary in the legal complex language [21]. In the meantime, the CNN unit derives local features and phrase-level patterns that are pertinent to legal classification.

• **Embedding Layer:** changes the token of the inputs into dense semantics vectors with respect to the already trained FastText embedding.
• **Bidirectional LSTM:** It not only has the forward contextual dependency in it, but also the backward contextual dependency.
• **1D Convolution:** local syntactic features such as legal phrasings and clause style.
• **Global Max Pooling:** Elimination of the dimensionality of features is done through the selection of the most vivid features.
• **Dense + SoftMax Layer:** A multi-class classification over the legal classes.

#### *B. Model 2: LSTM AND GRU*

The second architecture introduces a GRU instead of a CNN layer in order to make the computational process more efficient without having to force a trade-off between sequential learning. LSTM can handle relatively intricate temporal dependencies, and the GRU, being a simplified recurrent unit, speeds up training and serves as an aid in opposing overfitting.

• **Embedding Layer:** infuses an input sequence to already trained FastText vectors.
• **LSTM Layer:** Learning about the long-term dependency on the textual patterns in the law community.
• **GRU Layer:** It presents the effective recurrent layer that has fewer parameters than the LSTM. Dropout: A Small term was added to the regularization and prevent violating the issue of overfitting.
• **Dense + SoftMax Layer:** Here, it gives a likelihood value on every legal category

### 3.5. Models Training and Optimization

The training and optimization section entailed inputting the pre-processed legal textual data into the baseline model and into the hybrid deep learning model. The most important steps were fulfilled as follows:
• **Train-Test Split:** The data set was divided into the training, validation, and test sets, using the 80-10-10 split, to determine the robustness of the given model.
• **Loss Function and Optimizer:** In categorizing

the problem, we measure the loss using categorical cross-entropy. It was chosen because it makes learning rates adaptive and quickly converges.

• **Hyper-parameter Tuning:** These are the most significant hyperparameters that are learning rate, batch size, number of Epochs, and the rate of dropout, which were themselves tuned manually. Early stopping was done to avoid over-fitting, considering the validation loss.

• ***Regularization Algorithms:*** To assess the issue of an over fit, dropout (0.6) was implemented to conduct spontaneous disengagement of neurons in the course of training. L2 regularization and batch normalization were tried, but we decided not to use them because they did not help much.

## 4. RESULT AND ANALYSIS

This section provides a thorough examination of the testing results that were achieved by using classical machine learning models and deep learning architecture (LSTM+CNN and LSTM+GRU) for the problem of the classification of legal documents based on the EurLex dataset. The principal aim of the research was to determine the effectiveness of all models in making correct decisions pertaining to the classification of legislative documents, Directive, Regulation, or Decision, depending on the text material. The findings were assessed qualitatively (mainly visual assessment tools, confusion matrix, ROC curve, and others) and quantitatively (Accuracy, Precision, Recall, and F1-Score). The findings prove that hybrid system performance is much better than the conventional approach and provides better accuracy and resilience in processing complicated legal phrases. Some of the significant observations, a comparison of performances, and challenges are reported in the following sub-sections.

### *4.1. Results with Baseline ML models*

To initiate a baseline benchmark in the legal document classification task, we initially employed three traditional machine learning algorithms, namely, Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF). These models were chosen because they are popular and have already been tested in text classification issues. Features that were used to train each model were extracted using FastText word embedding, such that the average of its word vectors characterized each document. This way, we were able to encode the semantic information of the EurLex legal texts without jeopardizing the computational efficiency of the input representation. The results obtained from the traditional model are given in Table 2.

***Table 2: Machine Learning Model Evaluation Performance Metrics Scores***

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Logistic Regression | 95 | 0.94 | 0.95 | 0.94 |
| Support Vector Machine | 96 | 0.95 | 0.96 | 0.95 |
| Random Forest | 96 | 0.95 | 0.96 | 0.95 |

Logistic Regression was the most accurate amongst the baseline models, with an accuracy of 95%, showing that it could be effective in such linearly separable feature spaces. It was also computationally efficient and stable during training. The Support Vector Machine performed a bit better than the Logistic Regression at a rate of 96%. It performed well on high-dimensional input spaces, and thus it was well-suited to dense vector representations obtained with FastText embedding. On the same note, the Random Forest classifier managed to obtain 96 percent accuracy, as well as compared to the other models, which have the benefit of an ensemble learning paradigm to capture the non-linear association and complex decision lines. However, it took a slightly longer training time since it is a tree-based structure, and since it requires some guidance (tuning) of the parameters. Though these baseline models performed well, they have a very crucial limitation: they use static and averaged representations of text that fail to consider word order, syntactic structure, or contextual dependence. The use of subtle language and logical structuring is usually used in legal texts, and these models cannot capture them. Therefore, their effectiveness remains very powerful but is subject to inherent limitations since they cannot perform modelling on the sequential and hierarchical aspects of legal language.

### *4.2. Results with the Hybrid Deep Learning model*

To overcome the negative aspects of conventional machine learning models with respect to

sequential and contextual dependencies in legal texts, we applied two variants of hybrid deep learning models: LSTM+CNN and LSTM+GRU. The models utilize the strengths of recurrent layers (LSTM or GRU) to account for long-term dependencies and convolutional layers to account for local semantic patterns as well. All the models were trained on the FastText representations obtained from the EurLex set and are unchanged in terms of their maximum sequence length and dimension of embedding space. The performance that was measured is based on traditional classification measures: Accuracy, Precision, Recall, and F1-Score.

### *A. Result with Hybrid LSTM CNN*

The LSTM-CNN vision model architecture was applied, which was designed to recognize both orderliness and locality in the legal documents. It runs the FastText embedding and is evaluated on EurLex. The model achieved good results across all three categories of legal documents: Directives, Regulations, and Decisions. The results obtained from the Hybrid LSTM CNN are given in Table 3.

***Table 3: Summary of LSTM-CNN Model Evaluation Metrics***

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| LSTM+ CNN | 0.99 | 0.98 | 0.98 | 0.98 |

The Hybrid LSTM-CNN model achieved an overall accuracy of 99%, indicating a high level of performance in classifying legal documents into the three categories: Directive, Regulation, and Decision. Best among them was the Directive for which both precision and recall were nearly perfect, at 0.99, which showed that the model once again got consistent and discriminative features that are specific nature of this class.

### *B. Results with Hybrid LSTM-GRU Model*

The aim was to take the strengths of the Long Short-Term Memory (LSTM) networks that are able to capture long-range dependencies and mix them with the computational efficiency of Gated Recurrent Units (GRUs) to form the hybrid LSTM-GRU model. Like the previous model, it has been trained using a FastText representation and tested on the EurLex dataset. The results obtained from the Hybrid LSTM-GRU are given in Table 4.

***Table 4: Summary of LSTM-GRU Model Evaluation Metrics***

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| LSTM+ CNN | 0.98 | 0.97 | 0.97 | 0.97 |

The LSTM-GRU model also has an accuracy of 98% implying that it performed well in the classification of legal documents into their categories. Although this model is a bit lower than the LSTM-CNN one in the measures of precision and recall, the F1-score of 0.97 shows that it can produce decent equality in its rate of identifying legal document types effectively. This proves that this model is able to process linguistic complexity and semantic overlap, as it requires no convolutional layers.

### *4.3. Confusion Matrix and ROC Curve Visualization*

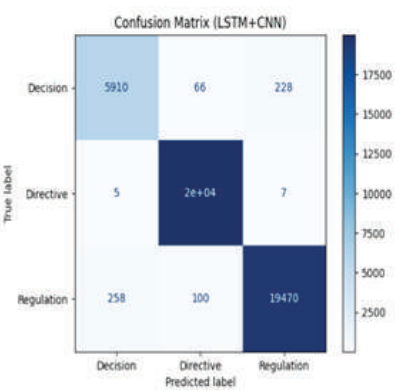



***Figure 2: Confusion Matrix of the LSTM +CNN***

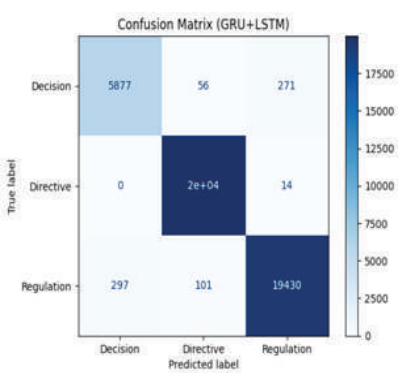



***Figure 3: Confusion Matrix of the GRU+LSTM model***

Confusion matrices of the two hybrid deep learning models were created to visualize the behavior of prediction and analyze them, as seen in Figure 2 (LSTM+CNN) and Figure 3 (GRU+LSTM). Such matrices are the illustrations of the optimum fitting that is rendered between projected and factual categories of all kinds of law.

Besides, ROC curves were computed with the help of a one-vs-rest scheme to assess classification thresholds and AUC (Area Under Curve) scores, which are presented in Figure 4 (LSTM+CNN) and Figure 5 (GRU+LSTM).

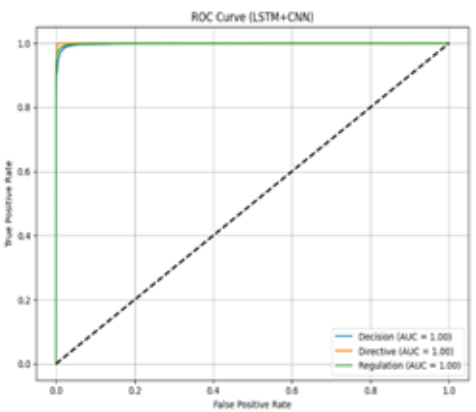



*Figure 5: ROC Curve (LSTM+CNN)*

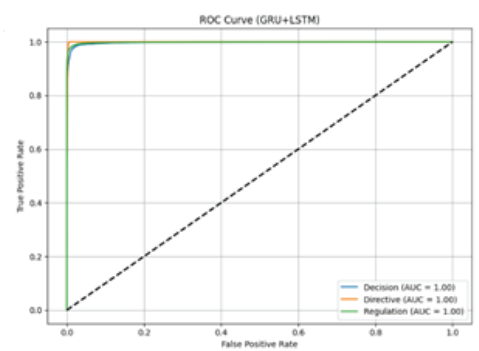



*Figure 6: ROC Curve (LSTM+GRU)*

The findings seem to show that the majority of the predictions are collated on the diagonal of the confusion map signifying great accuracy and minimum misclassification. Precisely, classification in Directive category was almost perfect whereas classification in Decision and Regulation types had a very less overlap. The findings were confirmed by ROC curve analysis whereby AUC was 1.00 in all the classes demonstrating strong generalization and very strong model performance.

### *4.4. Comparative Analysis and Key Observations*

In this section, in order to compare the performance of all the assessed models holistically, we contrast original examples of machine learning classifiers (Logistic Regression, SVM, Random Forest) with hybrid variants based on deep learning (LSTM+CNN and LSTM+GRU). The hybrid models greatly performed better than the traditional baselines when it comes to accuracy, precision, recall, and F1-score. Even though the baseline models were rather effective, with accuracy varying between 95 and 96 percent, they were hindered by the situation because they had no way of grasping the temporal and relational aspects of the legal speech.

- The network composed of LSTM+CNN has the highest accuracy of 99% as this model has the advantage of the LSTM to capture the sequential context and CNN to find the localized features.
- LSTM+GRU was not that far with an accuracy of 98, and is computationally efficient and converges faster owing to the simpler architecture of GRU.

***Table 5: Comparative Evaluation of All Models***

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Logistic Regression | 0.95 | 0.94 | 0.95 | 0.94 |
| Linear SVM | 0.96 | 0.95 | 0.96 | 0.95 |
| Random Forest | 0.96 | 0.95 | 0.96 | 0.95 |
| LSTM+ CNN | 0.99 | 0.98 | 0.98 | 0.98 |
| LSTM+ GRU | 0.98 | 0.98 | 0.97 | 0.98 |

Table 5 shows the comparative analysis of all the models applied to this study on the basis of four important performance measures, namely accuracy, precision, recall, and F1-score. Both Linear SVM and Random Forest reduced to the classical type of machine learning models, and both showed a high performance level, which was quite high, 0.96 accuracy. The second-best was the Logistic Regression having an accuracy of 0.95. Nevertheless, the hybrid deep learning models performed much better as compared to the

classical methodologies. LSTM+CNN model showed the best accuracy (0.99) and steady values of precision, recall, and F1-score (0.98 all of them). This shows that LSTM+CNN excels in the capacity to identify the semantics of the legal texts. The LSTM+GRU model also worked tremendously well with an accuracy of 0.98 which was still lower than the LSTM+CNN model, but still ahead of all the conventional techniques. These findings provide evidence as to why hybrid deep learning models are more useful in predicting legal decisions because they better manage complex and context-laden court language.

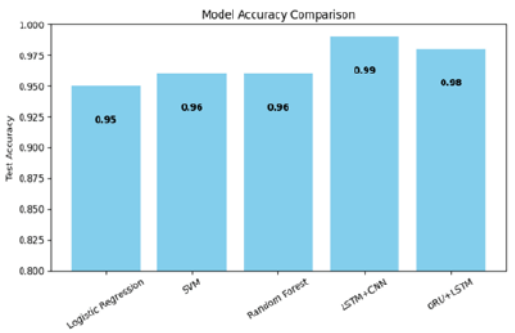



***Figure 7: Accuracy Comparison Chart***

To further exemplify the discrepancies in model accuracies, Figure 7 gives a bar chart comparison of all the five models. The max accuracy result was obtained by the hybrid (LSTM+CNN) with 99 percent, the closest result by LSTM+GRU at 98 percent matching the traditional machine learning-based models. This image represents the reliability of hybrid deep learning schemes to grasp complicated inferences in the legal sphere. The Comprehensive Performance Evaluation is demonstrated in figure 8. The graph presents a comparison of the solid performance of the all model on four essential measures of accuracy, precision, recall, and f1-score.

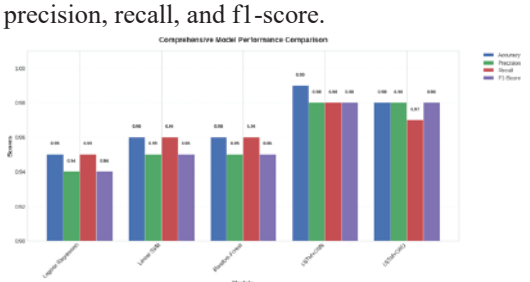



***Figure 8: Comprehensive Performance Evaluation***

### *4.5. Limitation*

Although the proposed hybrid deep learning architectures produced good results in categorizing the legal texts in the EurLex dataset, there exist a number of limitations. To start with, the research considerations were restricted to three of the legal categories: Directive, Regulation, and Decision, constraining the applicability of the findings to larger bodies of several legal areas, i.e., national laws, court cases, etc., and multilingual collections. Second, training such models as LSTM+CNN and GRU+LSTM is computationally expensive; therefore, such training requires GPU-capable environments and thus can be problematic otherwise, such as in low-resource or real-time applications. Third, even though DL models may have a good predictive accuracy, they are usually prone to problems with interpretability, not being able to clarify the reasons behind making some classification decisions, which is a serious shortcoming when it comes to legal and policy applications. Also, the models can be overfit to class imbalance in training and can perform poorly on inherently skewed or new ones without additional adjustment. Finally, FastText embeddings are significant because of the sub word-level benefits. Still, they cannot seem to grasp the legal-specific domain meaning, which implies that future research could involve domain-specific or contextual embeddings like LegalBERT to increase their quality.

## 5. CONCLUSION

The baseline machine learning models, as well as the multitasking deep learning models, were compared in this paper based on the accuracy with which they could be used to classify the legal texts found in the EurLex corpus. The resulting FastText embeddings and LSTM+CNN, accompanied by the GRU+LSTM model, made it possible for the proposed method could capture the semantic pattern and simultaneously capture the sequential pattern on legislative documents. The overall accuracy and macro F1-score obtained the best results in the LSTM +CNN model, hence considering it to be superior to the rest of the models regarding predicting the outcome of the legal cases. Whereas the simplicity in computation that the traditional models might portray, a deep understanding of the complexity of the legal language was better represented by the deep learning models. According to those findings, one could say that hybrid architectures can become a good tool that increases the appeal of automated legal text

classification, with implications that are worth considering in building intelligent legal information systems. However, the flaws related to computational cost and domain generalizability suggest that further research is necessary to continue working on the scalability aspect and extend the practice across various legal situations.

### *5.1. Future Work*

Future researchers will be able to consider the possibility of enlarging the dataset by touching on a larger variety of legal categories and multilingual corpora to enable a model to be more generalizable. It could be reasonable to use domain-specific embeddings like LegalBERT to enhance the legal semantic of law documents. More is recommended to be further carried out in matters of model interpretability where explainable AI ought to be applied to streamline MAML put into practice. In addition, to guarantee that the potentials of the proposed models are used in low-resource or realistic time legal applications, it is possible to create a lightweight or optimized version of the proposed models.

## REFERENCE


[1] J. Cui, X. Shen and S. Wen, "A Survey on Legal Judgment Prediction: Datasets, Metrics, Models and Challenges," *IEEE Access,* 11, pp. 102050–102071, 2023.

[2] N. Z. Dina, S. D. Ravana and N. Idris, "Legal Judgment Prediction using Natural Language Processing and Machine Learning Methods: A Systematic Literature Review," *SAGE Open,* 15(2), pp. 1–25, 2025.

[3] Y. Feng, C. Li and V. Ng, "Legal Judgment Prediction: A Survey of the State of the Art," *IJCAI Int. Jt. Conf. Artif. Intell.,* pp. 5461–5469, 2022.

[4] G. Sukanya and J. Priyadarshini, "Fine Tuned Hybrid Deep Learning Model for Effective Judgment Prediction Fine Tuned Hybrid Deep Learning Model for Effective Judgment Prediction," 142(3), March, 2025.

[5] M. H. Siddiqi et al., "A HybridFramework Combining Rule-Based and Deep Learning Approaches for Data-Driven Verdict Recommendations," *Comput. Mater. Contin.,* 83(3), pp. 5345–5371, 2025.

[6] G. Sukanya and J. Priyadarshini, "Fine Tuned Hybrid Deep Learning Model for Effective Judgment Prediction," *C. - Comput. Model. Eng. Sci.,* 142(3), pp. 2925–2958, 2025.

[7] D. Aumiller, A. Chouhan and M. Gertz, "EUR-Lex-Sum: A Multi- and Cross-lingual Dataset for Long-form Summarization in the Legal Domain," *Proc. 2022 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2022,* pp. 7626–7639, 2022.

[8] I. Chalkidis and N. Aletras, "Extreme Multi-Label Legal Text Classification: A case study in EU Legislation," pp. 78–87, 2019.

[9] M. Zheng, B. Liu, and L. Sun, "Internet of Things Era," 2022, 2023.

[10] J. Liu et al., "Legal Fact Prediction: The Missing Piece in Legal Judgment Prediction".

[11] D. Alghazzawi et al., "Efficient Prediction of Court Judgments Using an LSTM+CNN Neural Network Model with an Optimal Feature Set," *Mathematics,* 10(5), 2022.

[12] S. Feuerriegel et al., "Causal machine learning for predicting treatment outcomes," *Nat. Med.,* 30(4), pp. 958–968, 2024.

[13] Y. Wen and P. Ti, "A Study of Legal Judgment Prediction Based on Deep Learning Multi-Fusion Models—Data from China," S*AGE Open,* 14(3), pp. 1–10, 2024.

[14] A. P. Getman, O. M. Yaroshenko, R. V. Shapoval, R. Y. Prokopiev, and M. I. Demura, "The Impact of Artificial Intelligence on Legal Decision-Making," *Int. Comp. Jurisprud.,* 9(2), pp. 155–169, 2023.

[15] X. Shang, "A Computational Intelligence Model for Legal Prediction and Decision Support," *Comput. Intell. Neurosci.,* 2022(1), 2022.

[16] I. Chalkidis, M. Fergadiotis, and I. Androutsopoulos, "MultiEURLEX - A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer," *EMNLP 2021 - 2021 Conf. Empir.*

*Methods Nat. Lang. Process. Proc.,* pp. 6974–6996, 2021.

[17] M. Alali et al., "JUSTICE: A Benchmark Dataset for Supreme Court's Judgment Prediction," [Online], 2021.

[18] F. Ariai and G. Demartini, "Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges," *ACM Comput. Surv.,*1(1), [Online]. Available: http://arxiv.org/abs/2410.21306, 2024.

[19] M. A. Ibrahim, A. T. Handoyo and M. S. Anggreainy, "Hybrid Deep Learning for Legal Text Analysis: Predicting Punishment Durations in Indonesian Court Rulings," 2023.

[20] D. V. Carvalho, E. M. Pereira and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electron.,* 8(8), pp. 1–34, 2019.

[21] N. Bansal, A. Sharma, and R. K. Singh, "An evolving hybrid deep learning framework for legal document classification," *Ing. des Syst. d'Information,* 24(4), pp. 425–431, 2019.

[22] T. Santosh, M. H. Elganayni, S. Sójka and M. Grabmair, "Incorporating Precedents for Legal Judgement Prediction on European Court of Human Rights Cases," *EMNLP Conf. Empir. Methods Nat. Lang. Process. Find. EMNLP 2024,* pp. 3743–3750, 2024.

[23] E. Q. Nuranti, E. Yulianti, and H. S. Husin, "Predicting the Category and the Length of Punishment in Indonesian Courts Based on Previous Court Decision Documents," *Computers,* 11(6), pp. 1–20, 2022.

[24] S. Ramzan, M.M. Iqbal and T. Kalsum, "Text-to-image generation using deep learning," *Engineering Proceedings,* 20(1), pp. 16, 2022.

[25] S. Naseer et al., "Enhanced network anomaly detection based on deep neural networks," *IEEE access* 6, pp. 48231-48246, 2018.

[26] H. Tauqeer et al., "Cyberattacks detection in iomt using machine learning techniques," *Journal of Computing & Biomedical Informatics,* 4(01), pp. 13-20, 2022.

[27] T. Nazir et al., "EfficientPNet—an optimized and efficient deep learning approach for classifying disease of potato plant leaves," *Agriculture* ,13(4), pp. 841, 2023.