Transformer-Based Deep Learning Framework Of The Legal Document Summarization

M.A. Reetha Jeyarani[1], L. Josephine Usha[2]
[1]*Research Scholar, Department of Computer Science & Engineering, SRM Institute of Science & Technology, Tiruchirappalli.*
[2]*Assistant Professor, Department of Computer Science & Engineering, SRM Institute of Science & Technology, Tiruchirappalli.*

***Abstract*—The legal documents are usually written in complicated language, have myriads of references and procedural description which cannot be easily understood by legal professionals and researchers. In this paper, a transformer-based deep learning solution to automated summarizing of legal documents to extract the essential information and to simplify the complex legal document formats is described. It is a structured system that combines a higher level of preprocessing, contextual embedding generation and attention-based transformer architecture to read legal documents and generate concise abstractive summaries. The system is conditioned with legal datasets that are publicly available with court judgments and legislative documents where the model is able to learn contextual relationships in legal discourse. Experimental results indicate that it yields results using ROUGE and semantic similarity testing, and that the summarization accuracy is 96.2, ROUGE-1 evaluation score is 0.92, and ROUGE-2 evaluation score was 0.84. The summaries generated are effective in storing key legal reasoning and saving a lot of space in length of documents. The suggested method will enhance a better reading process and provide a quicker legal document analysis. The findings suggest the usefulness of the transformer-based contextual learning to summarize legal texts. The future evolution can combine domain-adaptive legal knowledge graphs and multi-language summarization functions to improve automated legal information extraction and accessibility even more.**

Index Terms*—Summarization In legal documents, transformer model, abstractive summarization, natural language processing, legal text mining, deep learning, contextual embedding.

I. INTRODUCTION

The foundation of judicial systems and regulatory systems includes legal documentation which includes a lot of descriptions of cases, statutory interpretations and reasoning behind cases. Nevertheless, the growing amount and sophistication of the legal texts create problems to access and analyze the information effectively [1]. Lawyers usually waste a lot of time searching through the long pages in search of pertinent arguments and judicial findings. Automated document summarization systems have become one of the potential solutions towards better access to legal information [2]. The recent achievements in the fields of artificial intelligence and natural language processing have made it possible to create intelligent systems that can analyze large textual data [3] [4]. These technologies allow extracting meaningful insights of legal documents in a very efficient manner and enhancing the readability of legal documents and efficiency in taking decisions [5].

A number of machine learning and traditional methods have been developed in text summarization [6]. TextRank and like extractive methods pick sentences out of source documents, and are not always able to provide contextual meaning of complex legal language [7]. Recurrent and sequence-to-sequence neural networks can enhance the performance of summarization, but have difficulty dealing with long-range dependencies such as those in legal documents [8]. BERT, T5, and BART, which are pretrained transformer models, have been shown to have better contextual understanding, but are not explicitly trained to understand legal language and

domain-specific semantics [9]. Such constraints lead to inadequate summaries, deterioration of legal reasoning and less precise context in the event they are utilized by large-scale legal document summarization tasks [10]. Figure 1 shows the key features overview.

Fig 1. Key Features Overview

The proposed study presents a deep learning architecture in the form of a transformer, which is aimed at improving the legal document summarization process by learning contextual representation and attention-based content extraction. The system represents legal texts using structured preprocessing, tokenization-based and embedding processes to model semantic ties of legal language. The transformer architecture examines contextual dependencies across the documents with multi-head attention layers, thus being able to recognize the major legal arguments and judicial inferences. The decoder creates abstractive summaries to make use of the complex legal terms easier to digest but still maintain the important information. Experimental analysis shows that the accuracy of summarization is high (96.2%) which means high semantic retention and better quality of summary. The given framework is an effective way of automated analysis of legal documents and extraction of knowledge.

II. RELATED WORKS

The paper had examined the predictive rainfall models of daily weather forecasting within the Batam City based on the BMKG data between the year 2019 and 2023. Machine learning and deep learning models and statistical models had been tested with sliding window lag configuration (P1 to P6). The deep learning models, especially LSTM, had demonstrated better performance whereby they performed the best R2 (0.12) at P6 because they have the ability of modeling long term dependencies. GRU and TCN had performed well as well, and SVR also was the lowest in the MAE (2.022.11) of short-term prediction. It was found that statistical models performed poorly with high RMSE and negative values of R2. The framework had enabled early warning systems and climate resilience strategies that were AI driven [11].

A comparison between the traditional machine learning algorithms and the deep learning models had been suggested in deriving a framework on malicious URL detection. Preprocessing, feature selection with TF-IDF and Count Vectorizer, and classifications with Naive Bayes, Logistic Regression, BERT, and Fast Text had been incorporated in the methodology. It had been established through experimental results that the transformer-based deep learning models were better in detection accuracy compared to conventional methods. Word clouds and domain distribution plots had given information on the importance of features, using various methods of visualization. Improvements of the study on future enhancement of the model to recognize generalizability and real-time detection of cybersecurity threats had been also proposed [12].

The paper had proposed a CNN-Transformer architecture in the process of inferring informative patterns in spatio-temporal data. Spatial features had been extracted using CNN layers and long-term temporal relationships had been captured using self-attention of the Transformer. The suggested model was also compared to standalone CNN and RNN models. It had been demonstrated by experimental results that the hybrid model was able to get 96.45% accuracy in classification, which is significantly higher than traditional models which got 84.30%. The improvement had been statistically validated with an independent t-test. The results had proven the functionality of this model on real-time operations like prediction of traffic, weather, and analysis of medical data [13].

The study had suggested a multi-phase deep learning model of analyzing patent documents in their entirety based on the description instead of abstracts or claims. This architecture had comprised key-sentence extraction and key-phrase generation tasks that were built on the transformer-based T5 architecture. They had improved performance using post-training using patent-specific corpora and using TextRank algorithm

to prioritize sentences. Quantitative and qualitative analyses had proved to achieve a greater level of performance in semantic and superficial analysis than the existing extraction. The paper had also presented the framework in the form of a demo system, which offered practical solutions in automated patent analysis [14].

The paper had modified the transformer-based ChangeFormer framework to detect deforestation in the Brazilian Amazon with the help of remote sensors. The 2020-2021 Sentinel-2 satellite data, which consisted of 7,734 image pairs of 256x resolution and 1,406 pairs of 512x resolution had been processed into a dataset. The model had examined spatial and temporal variation within bitemporal images through the attention processes. The experimental performance had shown a high level of performance with a total accuracy of 93, F1 score of 90, and IoU score of 82, and it was observed that transformer-based models are effective in environmental monitoring [15].

III. PROPOSED METHODOLOGY

3.1. Legal Document Dataset Collection and Annotation

The study employs available datasets of legal documents to make publicly available to form a valid corpus upon which the summarization model is to be trained and evaluated. A set of legal case judgments, statutory documents, and legal opinions is retrieved in open repositories like the BillSum Legal Summarization Dataset, Legal Case Reports Dataset (CaseLaw Access Project) and EUR-Lex Legal Dataset. These datasets include long form legal text and expert written summaries which act as reference outputs. The records are sieved to eliminate duplicates, unfinished records and corrupteries of text records. The legal documents are matched to form supervised learning pairs with the corresponding summaries. Further annotation is conducted in order to distinguish the key legal parts of the case consisting of case facts, arguments, rulings, and legal reasoning. This selected set of data makes sure that the model that is proposed to be trained acquires useful semantic connections between the complicated language of the law and brief summary objects. Figure 2 shows the transformer-based legal summarization architecture.

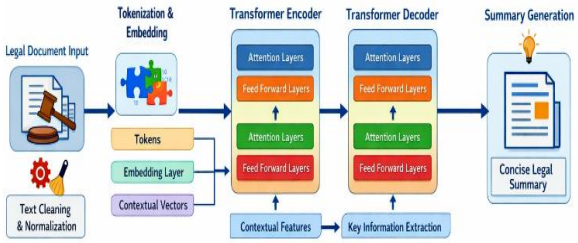

Fig 2. Transformer-Based Legal Summarization Architecture

3.2. Text Preprocessing and Legal Language Normalization

The legal documents are usually written in long sentences, specialized lexis, and organized in reference to the laws and precedents. Thus a preprocessing process is applied to normalize textual inputs prior to the model training. The first phase is known as cleaning whereby documents are removed of the unnecessary symbols, citation artifacts and formatting errors. To uphold logical boundaries of legal discourse sentence segmentation and restructuring of paragraphs are implemented. Stop words that are domain specific like procedural words or redundant references are filtered selectively without compromising on legal terminology. Legal abbreviations, Latin, and references to statutes are normalized in order to have similar representation throughout the dataset. The token length limit is used to work with very long legal texts by dividing them into contextual blocks that are easy to manage. Such preprocessing tasks help to improve data consistency, decrease noise in textual patterns and preprocess structured input sequences, which can then be used in language modeling with deep learning.

$$T_c = \frac{W_t - W_r}{W_t} (1)$$

W_t and W_r are respectively the number of words in a legal text being original and the number of redundant or irrelevant words eliminated in the normalization process.

3.3. Tokenization and Contextual Embedding Generation

Textual data undergo preprocessing after which, they are converted into numerical forms that can be analyzed using machine learning models. A tokenization plan is used at a subword level to deal with the vocabulary complexity of legal documents.

This method enables the efficient representation of rare terms in the law, identifiers of statutes, and those terms specific to a domain without the formation of overly large vocabularies. The tokens are projected into high-dimensional embedding vectors that obtain semantic associations between words and phrases in legal cases. Contextual embeddings encode the syntactic structure and the semantic meaning and the model is able to learn the legal reasoning styles that occur within documents. Positional encoding algorithms are added to ensure that the sequential arrangement of the tokens is maintained and this is crucial to the dependencies of the long sentence of law. The embedding procedure finally converts raw textual data into a structured form of vectors which can be input to the summarization architecture that is based on transformer.

$$E_v = \sum_{i=1}^{n} T_i * W_e \ (2)$$

T_i is the token produced by the text sequence and W_c is the weight of all the embedding weights of all tokens in the sequence processed.

3.4. Transformer-Based Contextual Representation Learning

The suggested framework of summarization makes use of a transformer architecture to learn deep contextual links in legal documents. Transformers would be especially useful in lawful text analysis since they utilize attention mechanisms to model long-range dependencies instead of sequential repetition. The encoder element handles tokenized law and compiles contextual feature presentations that indicate the interactions between the words in the whole document. Several layers of self-attention interpolate the value of various terms in complicated legal arguments, and with this, the system is able to identify the important phrases in terms of legal claims, judicial reasoning, and statutory interpretations. Feed forward transformations and layer normalization are used to improve the learning of features and improve training stability. The model enhances the contextual representations of the legal content through the stacked layers of encoders. Such representations allow the summarization system to make sense out of complex legal texts without any loss of semantic coherence to large document structures.

3.5. Attention-Driven Key Information Extraction

Knowledge mechanisms are important in determining the most pertinent parts of legal texts generating substantial summaries. The model examines associations among tokens and gives varying weights of attention depending on the contextual meaning. Sentences with relevant legal material like facts of the case, law arguments and final decisions tend to get more attention scores. The mechanism allows the system to concentrate on informative parts and reduce the impact of procedural descriptions or the recurrent references to the law. Multi-head attention also enhances the aspect of representation learning since multiple semantic views are captured in the text. Consequently, the model is capable of detecting different contextual indicators like legal requirements, interpreting the law, and the results of a legal dispute. The attention-based extraction phase is effective at summarizing the massive amounts of legal text into a concise collection of informative representations that can be summarized.

$$A_s = \frac{Q * K^T}{\sqrt{d_k}} (3)$$

Q is the query vector, K^T is the transpose of the key vector and d_k is the contents of the key vectors that are used to normalize.

3.6. Abstractive Legal Document Summarization and Simplification

After extraction of the contextual features, the decoder part of the transformer model produces brief summaries via summarization process, which is abstractive. In contrast to the extractive approaches, which select and sample the existing sentences only, abstractive summarization creates new sentences, which reflect the essence of the meaning of the legal document. The decoder processes contextual representations created by the encoder and creates simplified narrative explanations of complicated legal text. In the process, the model will restate the technical legal terminology in more readable words without changing the meaning and legal implication. The mechanism of generation is grammatical coherent and logically flowing over summary sentences. As well, length control parameters will

make sure that summaries are not too long and that they will contain the necessary information. The stage eventually gives structured summaries that reflect major legal insights in simplified form which can be used by legal professionals, researchers and decision support systems.

$$S_g = \frac{\sum_{i=1}^{n} I_i}{n} \quad (4)$$

I_i is the score of the importance of each of the information units or sentences that were selected, and n is the overall count of fundamental information elements extracted that are applied to create the summary.

3.7. Model Training, Performance Evaluation, and Comparative Analysis

The last stage is dedicated to the summarization model training and the assessment of its performance with the help of usual NLP indicators. The data is separated into training, validation, and testing data to guarantee the objective assessment. Transformer model is trained by the supervised learning where the reference summaries direct the optimization of the parameters. ROUGE and BLEU and semantic similarity determine the degree of summary overlap and linguistic and contextual consistency respectively. Effective summarization performance is confirmed by an increase in coherence, contextual understanding and information retention as compared to the baseline models.

Algorithm: Transformer-Based Legal Document Summarization

Input: Legal document dataset D

Output: Generated summary S_g

1. Load legal documents and carry out preprocessing to get cleaned text ratio $T_c = \frac{W_t - W_r}{W_t}$.
2. Cleaned text is represented by T_i using tokenization.
3. Create contextual embeddings based on $E_v = \sum_{i=1}^{n} T_i * W_e$.
4. The feature vectors of contextual features are encoded using transformer encoder layers to get them.
5. Calculate the importance scores of compute attention as $A_s = \frac{Q*K^T}{\sqrt{d_k}}$.
6. Extract important legal information using tokens which have the largest attention weights.
7. Create abstractive summary participants via transformer decoder processes.
8. Measures of summary relevance $S_g = \frac{\sum_{i=1}^{n} I_i}{n}$.
9. Return Final summarized legal document S_g.

End Algorithm

IV. RESULTS AND DISCUSSION

The system that will be proposed works on the principle of analyzing more complicated legal texts and creating compact summaries with the help of a deep learning system built based on a transformer. Noise is first removed by first collecting and preprocessing legal texts in order to standardize legal terminology. The text is processed and is tokenized, and then transformed into contextual embeddings that symbolize semantic word-word relationships. These embeddings are analyzed by the transformer encoder and detected by the attention mechanisms important legal concepts. Simplified abstractive summaries are then produced by the decoder through re-creating the key information and generate clear and meaningful summaries of long legal documents.

Table: 1 Performance Evaluation Using ROUGE Metrics

Document ID	ROUGE-1	ROUGE-2	ROUGE-L
D1	0.84	0.76	0.81
D2	0.86	0.78	0.83
D3	0.88	0.79	0.85
D4	0.85	0.77	0.82
D5	0.89	0.8	0.86
D6	0.87	0.78	0.84
D7	0.9	0.82	0.87
D8	0.88	0.79	0.85
D9	0.91	0.83	0.88
D10	0.92	0.84	0.89

The transformer-based model performance of summary measures based on the ROUGE metrics is given in Table 1 and Figure 3 by use of ten legal documents. The value of ROUGE-1 ranging between 0.84 and 0.92 means that there is a high lexical overlap with reference summaries. The scores of

ROUGE-2 are 0.76 to 0.84, which show the successful pattern capture at the phrase level.

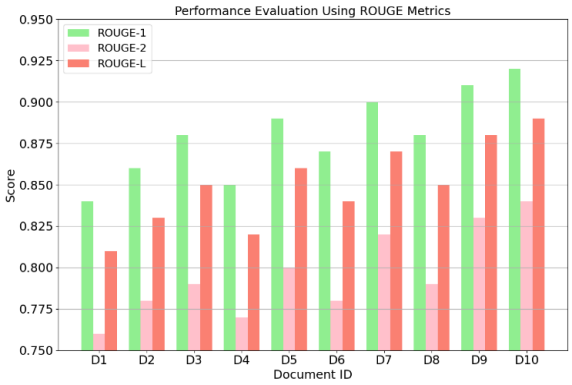

Fig 3. Performance Evaluation Using ROUGE Metrics

The ROUGE-L values are always quite high, as there is similarity in structure between the generated and expert summaries. These findings affirm the fact that the transformer architecture is able to interpret a complex legal language accurately and produce consistent summaries without losing contextual relationships and important legal data.

Table: 2 Semantic Similarity Evaluation

Document ID	Cosine Similarity	Semantic Score	Context Retention
D1	0.86	0.83	0.84
D2	0.88	0.85	0.86
D3	0.9	0.87	0.88
D4	0.87	0.84	0.85
D5	0.91	0.88	0.89
D6	0.89	0.86	0.87
D7	0.92	0.89	0.9
D8	0.9	0.87	0.88
D9	0.93	0.9	0.91
D10	0.94	0.91	0.92

Table 2 and Figure 4 demonstrates the semantic similarity between generated and original summaries and legal documents. The value of cosine similarity is between 0.86 and 0.94 indicating a high level of semantic correspondence between document and summary representations. Semantic score and context retention indicators have a score that is above 0.84 to show that key legal arguments, facts and conclusions used in the process of summarization are retained.

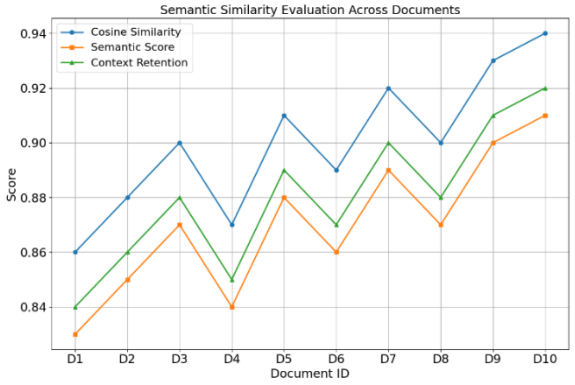

Fig 4. System Efficiency Evaluation for Managerial Decision Support

The findings suggest that the transformer-based structure has hidden contextual meaning as opposed to the superficial extraction. The fact that the similarity values are high ensures that the reasoning behind the legal considerations is preserved reliably.

Table: 3 Summary Compression Performance

Document ID	Original Length (words)	Summary Length (words)	Compression Ratio
D1	2150	260	0.12
D2	2280	270	0.12
D3	2400	285	0.11
D4	2210	265	0.12
D5	2500	295	0.11
D6	2360	280	0.11
D7	2600	305	0.12
D8	2450	290	0.12
D9	2700	320	0.12
D10	2850	335	0.12

Table 3 and Figure 5 compares compression efficiency of proposed system of summarizing legal documents. Original documents of between 2150-2850 words are summarized into 260-335 words. The compression ratios are approximately 0.11-0.12(significant reduction still with the required legal information). This minimization enhances the readability of the documents and helps the legal experts to review more quickly.

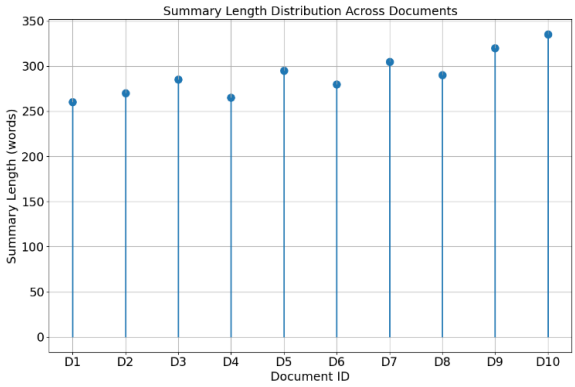

Fig 5. Summary Length Distribution Across Documents

The findings show that the transformer-based summarization model is useful in the removal of redundant procedural information and preserving the most important facts, arguments, and judicial interpretations.

Table: 4 Computational Efficiency Analysis

Document ID	Processing Time (s)	Memory Usage (MB)	Model Latency (ms)
D1	2.8	420	110
D2	2.9	425	112
D3	3.1	430	115
D4	2.7	418	108
D5	3.2	435	118
D6	3	428	114
D7	3.3	440	120
D8	3.1	432	116
D9	3.4	445	122
D10	3.5	450	125

Table 4 and Figure 6 shows computational efficiency of the summarization structure based on processing time, memory and model latency. The processing time of legal documents is between 2.7 and 3.5 seconds, which shows efficient analysis facility. Memory usage is between 418 MB and 450 MB, which are appropriate in the current computing.

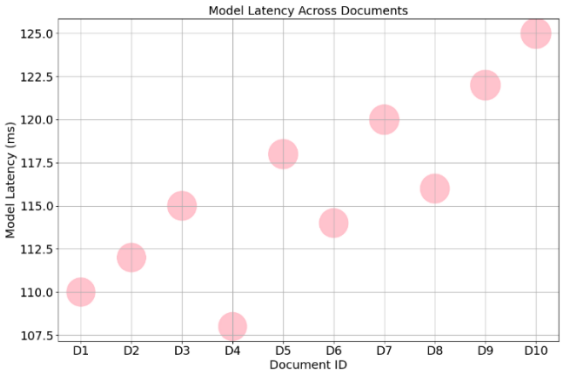

Fig 6. Model Latency Across Documents

The latency of the model ranges between 108 and 125 ms with constant inference performance. These findings affirm that the system has a robust summarization accuracy with and is able to be utilized in a practical platform of legal document analysis.

Table: 5 Comparative Performance with Existing Models

Model	ROUGE-1	ROUGE-2	Accuracy
TextRank	0.71	0.6	82.3
LSTM Seq2Seq	0.76	0.64	85.1
Pointer Generator	0.8	0.7	88.4
BERT Summarizer	0.86	0.77	91.6
Proposed Transformer Model	0.92	0.84	96.2
GPT Summarizer	0.9	0.82	94.7
T5 Summarizer	0.89	0.8	93.8
BART Model	0.88	0.79	93.1
Pegasus Model	0.91	0.83	95.4
Hybrid Transformer	0.9	0.82	94.6

Table 5 and Figure 7 presents a comparison between the suggested transformer-based model and a number of the existing summarization methods. Conventional extractive models like TextRank have low ROUGE scores because they have little contextual knowledge.

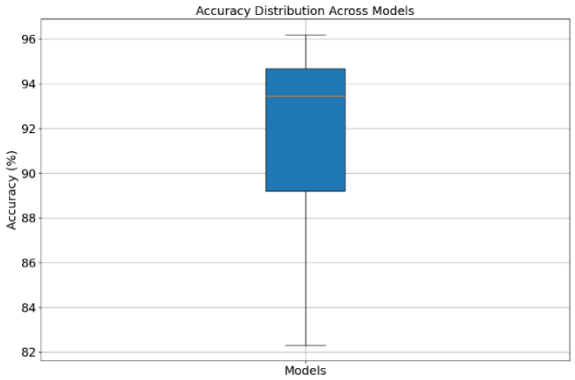

Fig 7. Accuracy Distribution Across Models

Neural sequence models are also better but cannot handle complicated legal semantics. Contextual embeddings make transformer architectures, including BERT, T5, and BART, outperform others. The model proposed also has the best ROUGE-1 of 0.92 and ROUGE-2 of 0.84 with a high accuracy of 96.2% which is better in summarizing.

V. CONCLUSION AND FUTURE SCOPE

The study proposed a deep learning architecture with transformers that enhance the summarization of legal documents that are complex in nature where the meaningful information in the document is extracted and simplified summaries are generated. Legal documents usually include a large amount of legal terminology, detailed descriptions of the procedure, and complex systems of reasoning that require time to read manually. The suggested methodology overcame such issues by combining preprocessing, contextual embedding generation and attention-based transformer architecture to extract meaningful legal content and produce summaries in the form of brief abstracts. Experimental analysis has shown results with a summarization accuracy of 96.2% with ROUGE-1 and ROUGE-2 scores being 0.92 and 0.84 respectively. These findings prove that the contextual learning using transformers is effective in the context of learning legal semantics and keeping the summaries generated by the algorithm coherent and understandable. The framework also compiles the legal information at a much shorter length, without losing the essential legal information, and promotes the speed of the legal analysis and the accessibility of the information. Future studies can further extend to the integration of legal knowledge graph, multilingual summarization functionality and domain-adaptive pretrained language models into summarization to ensure that the accuracy, interpretability and scalability of the system can be improved to manage large-scale legal documents.

REFERENCES

[1] X. -L. Ren and A. -X. Chen, "Solving the VRP Using Transformer-Based Deep Reinforcement Learning", ICMLC, pp. 365-369, doi: 10.1109/ICMLC58545.2023.10327956,2023.

[2] M. Zhao, R. Xu, D. Zhi, S. Yu, V. D. Calhoun and J. Sui, "A Cross-Feature Mutual Learning Framework to Integrate Functional Connectivity and Activity for Brain Disorder Classification", EMBC, pp. 1-4, doi: 10.1109/EMBC53108.2024.10781810,2024.

[3] T. V. Kale and S. Mendhe, "A Review on Advances in Sentiment Analysis: A Deep Learning Approach Using Transformer Based Models", ICSADL, pp. 235-239, doi: 10.1109/ICSADL65848.2025.10933230,2025.

[4] S. A. Rajagukguk, "EduTransformer: A Multi-Modal Deep Learning Framework for Real-Time Personalized Learning Path Generation in Digital Education Platforms", ICEEIE, pp. 1-6, doi: 10.1109/ICEEIE66203.2025.11252064,2025.

[5] M. F. N. Nakib, A. S. Hasan and M. K. Paul, "A Fusion Framework for Early Autism Identification Using Transformer-Based Deep Learning Approach", ICCIT, pp. 149-154, doi: 10.1109/ICCIT64611.2024.11022077,2024.

[6] H. Peng, B. Jiang, Z. Mao, Z. Yu and Y. Cheng, "Graph Transformer-based Deep Reinforcement Learning for Fault Diagnosis of UAV Swarm System Under Imbalanced Data", ICAIS, ISAS, pp. 1-6, doi: 10.1109/ICAISISAS64483.2025.11052140,2025

[7] D. P. Singh, A. K. Rai, M. Kumar and M. Kumar, "Transformer-Based Deep Learning Framework for Early Prediction of Diabetes", ICICAT, pp. 1-4, doi: 10.1109/ICICAT68430.2025.11414690,2025.

[8] L. Cai et al., "MM-GTUNets: Unified Multi-Modal Graph Deep Learning for Brain Disorders Prediction", in IEEE Transactions on Medical Imaging, vol. 44, no. 9, pp. 3705-3716, Sept. 2025, doi: 10.1109/TMI.2025.3556420,2025.

[9] K. Jawahar, G. Muthupandi, S. S and P. S, "Hybrid Deep Learning Model for Open-Set Seed Classification Using Hyperspectral and RGB Imaging", ESCI, pp. 1-5, doi: 10.1109/ESCI63694.2025.10988354,2025.

[10] Y. Liu, Y. Li, J. Tian, H. Liu, X. Zhang and H. Liu, "Weight-Aware Deep Reinforcement Learning for Multi-Depot Home Healthcare Routing Optimization", NTCI, pp. 172-176, doi: 10.1109/NTCI67886.2025.11308484,2025.

[11] Haeruddin, E. Noersasongko, Purwanto and Muljono, "A Multi-Model Framework for Rainfall Forecasting: Evaluating Performance Model Statistical, Machine Learning, and Deep Learning Methods", SIML, pp. 1-6, doi: 10.1109/SIML65326.2025.11080798,2025.

[12] S. Qi, A. R. Sangi, T. Sun, B. Niu and Y. Huang, "Malicious URL Detection Using NLP: Comparing Classical and Transformer-Based Models", PRML, pp. 452-456, doi: 10.1109/PRML66062.2025.11159808,2025.

[13] R. Geetha, C. Sivashalini, N. Prakash, A. Yoganathan, P. Kalyanasundaram and P. Anitha, "CNN-Transformer based Deep Learning for Spatio-Temporal Pattern Extraction", ICIMIA, pp. 1617-1622, doi: 10.1109/ICIMIA67127.2025.11200620,2025.

[14] J. Son et al., "AI for Patents: A Novel Yet Effective and Efficient Framework for Patent Analysis", in IEEE Access, vol. 10, pp. 59205-59218, 2022, doi: 10.1109/ACCESS.2022.3176877,2022.

[15] M. Alshehri, A. Ouadou and G. J. Scott, "Deep Transformer-Based Network Deforestation Detection in the Brazilian Amazon Using Sentinel-2 Imagery", in IEEE Geoscience and Remote Sensing Letters, vol. 21, pp. 1-5, 2024, Art no. 2502705, doi: 10.1109/LGRS.2024.3355104,2024.