

A Signal-Based Comparative Analysis Of Transformer And Cnn Architectures For Natural Language Processing Applications

Jaiprakash Narain Dwivedi^{1*}, Kailash Kumar Pareek², Ch V S Satyamurty³, Sanjay Madaan⁴, Gadde Mamatha⁵, K. Vigneshwari⁶

^{1*}Department of Information Technology, Parul Institute of Engineering and Technology, Faculty of Engineering and Technology, Parul University, Vadodara, Gujarat, India, Email: jpnndwivedi@gmail.com

²Department of Information Technology, Parul Institute of Engineering and Technology, Faculty of Engineering and Technology, Parul University, Vadodara, Gujarat, India, Email: kailash.87pareek@gmail.com

³Associate Professor, CVR College of Engineering, Department of Information Technology, Orcid Id: 0000-0002-0456-9742, Email Id: satyamurty@cvr.ac.in

⁴Department of Computer Engineering and Applications, GLA University, Mathura, Uttar Pradesh, India, Email Id: sanjaymadaan87@gmail.com, Orcid Id: <https://orcid.org/0000-0002-3583-0552>

⁵Assistant Professor, Department of CSE, Koneru lakshmaiah education foundation, Email id : g.mamatha@klh.edu.in, Orcid Id: 0009-0002-2195-6602

⁶Department of AIML, IFET College of Engineering, Email Id: vigneshwari0397@gmail.com writ fb 2

Abstract

Transformer-based networks and convolutional neural networks are popular sequence modeling models in natural language processing and have not been fully studied in terms of their resistance to various linguistic perturbations. This study presents a comparative analysis of CNN and Transformer models using a diagnostic hate speech detection benchmark comprising 3,728 linguistically controlled test cases. A character-level CNN and a token-based Transformer were evaluated under identical experimental conditions. The Transformer achieved higher overall performance, with an accuracy of 76.9% and an F1-score of 0.80, while the CNN achieved an accuracy of 71.4% and an F1-score of 0.75. The Transformer also exhibited a lower false negative rate of 19.6%, whereas the CNN showed a false negative rate of 28.4%. Functional analysis indicated that the Transformer performed better on semantically complex constructions such as negation and quotation. In contrast, the CNN demonstrated greater robustness to spelling-based perturbations, attaining accuracy values of up to 84.6% under character-level noise. These findings indicate that architectural effectiveness depends on the dominant characteristics of the linguistic input and the nature of the perturbation. The results provide quantitative evidence to support informed architecture selection for NLP applications requiring robustness to linguistic variability.

Keywords: Transformers, convolutional neural networks, natural language processing, robustness, hate speech detection.

1. Introduction

With multiple applications in domains including computer vision, speech processing, and natural language processing (NLP), deep learning has evolved into a paradigm of modern signal and information processing to allow tremendous progress in these fields. The convolutional neural networks (CNNs) and the Transformer-based models are two fundamentally dissimilar models of sequence modeling among deep learning networks. CNNs are good at localization of receptive fields and hierarchical feature extraction, which is why they are effective at describing local patterns and robustness to noise representations, whereas Transformers apply self-attention layers in order to describe global relationships between complete sequences [1]. Convolution-based models to attention-based models have been found in several aspects of use. Transformers have shown to be better at long-range spatial dependencies in visual perception and object detection, but CNNs are also competitive in terms of efficiency and robustness [2]. Hybrid architectures with CNNs and Transformers also imply that these two architectures complement each other in terms of signal representation and do not provide direct alternatives [3]. Industrial fault detection has also been reported to follow similar architecture trends with

CNN-based Transformer models, improving the sensitivity to signal complex behavior [4]. Transformers have become the predominant architecture in NLP since they can learn contextual semantics when dealing with very long sequences of text. The transformer-based language models have already achieved state-of-the-art in a large variety of NLP tasks, such as text classification and sequence labeling [5]. Their application in practical systems and toolkits has also increased their use in NLP systems [6].

Transformer architectures have a long history of being studied in the context of NLP, being noted to be scalable and powerful in their representations. Extensive research highlights the fact that self-attention allows modeling semantic relationships that are challenging to learn with the conventional recurrent or convolutional models [7]. Recent reviews of NLP techniques confirm the popularization of Transformers with the observation of sensitivity of input representation and noise [8]. In addition to NLP, Transformers have been used successfully on vision-based and multimodal signal analysis problems. Applications like road accident detection can be used to illustrate how attention mechanisms can be useful in processing complex visual signals that are both in time and space [9]. Transformer-based models have been demonstrated to be useful in speech and emotion recognition, where they demonstrated superior results in long-term dependency capture in sequential data [10]. Regardless of such advances, CNNs remain an important element in signal processing applications. Observations about the analysis of physiological signals indicate that CNNs are still very efficient when the signals are decoded as structured representations, especially in the case of noisy or resource-limited environments [11]. The electromyography and gesture recognition research also suggests that the accuracy of classification is highly contingent upon the strategy of feature extraction and the nature of signals, where Transformers handle globally structured signals and CNNs provide tolerance to localized distortions [12]. Similar comparative studies in acoustic emission-based corrosion sensors also indicate that Transformer models that have been used have been found to be superior to traditional methods in situations where signals have intricate time-dependence [13]. Despite an extensive research background on CNNs and Transformers in both the NLP and signal processing fields, the literature currently available is skewed in discussing aggregate performance metrics or hybrid architectural design. It lacks a systematic comparative analysis of these architectures in a signal-oriented manner that explicitly addresses the responses of these architectures to varying kinds of signal perturbation in NLP. Specifically, not many studies explore the comparative efficiency of surface-level distortions, i.e. in spelling differences or character-level noise, to semantically complex constructions, i.e. negation, quote, or counter-speech in affecting model behavior. Furthermore, textual data are hardly ever processed in terms of signal-processing, when it is obvious that linguistic distortions have the same similarity with noise in conventional signal processing. This gap restricts knowledge of the robustness of architecture and the error behavior of real-world NLP systems, particularly in safety-critical settings, like hate speech detection.

This paper aims to conduct a comparative signal-level analysis of CNN-based and Transformer models in natural language processing with controlled linguistic perturbations. The comparison of overall classification performance, the analysis of the architectural behavior of the linguistic representations they consider, both surface-level and semantically complex, the ability to resist noise at the character level, and the traits of errors by class are analyzed. The analysis of textual input as a linguistic signal that can be transformed in controlled ways will help in offering an ordered assessment of architectural appropriateness to NLP applications that shall be resilient to linguistic change.

2. Methodology

2.1 Research Design and Data Source

The research design was a comparative experimental study to examine the performance features of convolutional neural network (CNN) and Transformer-based architectures in the controlled linguistic signal transformations. The experiment aims at analyzing robustness, generalization of the functions, and the manner in which an error is handled by isolating the effects of the architecture as the training and evaluation conditions remain constant. The HateCheck benchmark that was used as an experimental dataset was a diagnostic dataset that was developed to test the hate speech detection models under linguistically controlled conditions [14]. The data is a collection of 3,728 instances of English-language texts that are labeled as either hateful or non-hateful and are grouped into prescribed functional categories that address a specific linguistic phenomenon. Using the dataset was simply utilized as a benchmark of evaluation to examine model behavior in controlled linguistic signal perturbations

2.2 Functional Signal Categories

The functional categories were classified on the basis of their overriding linguistic features. Superstratum categories are surface changes in spelling and use of explicit slurs, which are mostly lexical on local patterns. Negation, referred or quoted hate, and counter-speech are semantically complicated categories, and they need to be interpreted contextually instead of surface-level. The division helps to allow a systematic study of architectural behavior through different signal transformations.

2.3 Model Architectures

There were two neural architectures tested. Convolutional neural network (CNN) processes character-level input representations and uses several one-dimensional convolutional filters of different kernel widths, followed by max-pooling and a fully connected classification layer. This layout is designed with a focus on feature extraction in the locality and resilience to character-level distortions. The transformer model is composed of an encoder, which is based on self-attention and acts upon tokenized input sequence. They use positional embedding to maintain sequence order and pooled sequence embedding in order to aggregate the sequence representations to classify. The architecture aims at capturing the long-range dependencies and compositional semantics. Both models also had similar parameter scales, so that the performance difference found is mainly a result of architecture.

2.4 Training Procedure

The two models were developed based on a cross-entropy loss function and were optimized with the Adam optimizer. The same training conditions, such as batch size, learning rate and stopping criteria, were used across architectures. Hyperparameters were chosen in order to ensure stable convergence.

2.5 Evaluation Metrics and Error Analysis

Accuracy, precision, recall and F1-score were used to assess model performance. It was further analyzed in terms of performance based on functional signal categories to determine strength during certain linguistic changes. Character-level noise was tested for robustness by isolating spelling-based perturbation classes. False positive and false negative rates were assessed on the basis of confusion matrices, which were used to measure error behavior.

3. Results

3.1 Overall Classification Performance

The original experiments check the international effectiveness of the two structures with the entire HateCheck data. Accuracy, precision, recall and F1-score are used in reporting performance to adjust the imbalance in classes.

Table 1. Overall Classification Performance on HateCheck

Model	Accuracy (%)	Precision	Recall	F1-score
CNN	71.4	0.73	0.78	0.75
Transformer	76.9	0.77	0.83	0.80

The transformer model produces a better result than CNN in terms of all three metrics, being 5.5% more accurate and having a better F1-score (Table 1). The increase in the recall and sensitivity means a higher capacity to recognize hateful content correctly, which can be credited to the contextual modeling of the Transformer and the long-range dependency of the contextual modeling. The change trends of the performance of the two are presented in Figure 1(A) and 1(B), and the independent analysis indicates the stable superiority of the Transformer according to aggregate indicators.

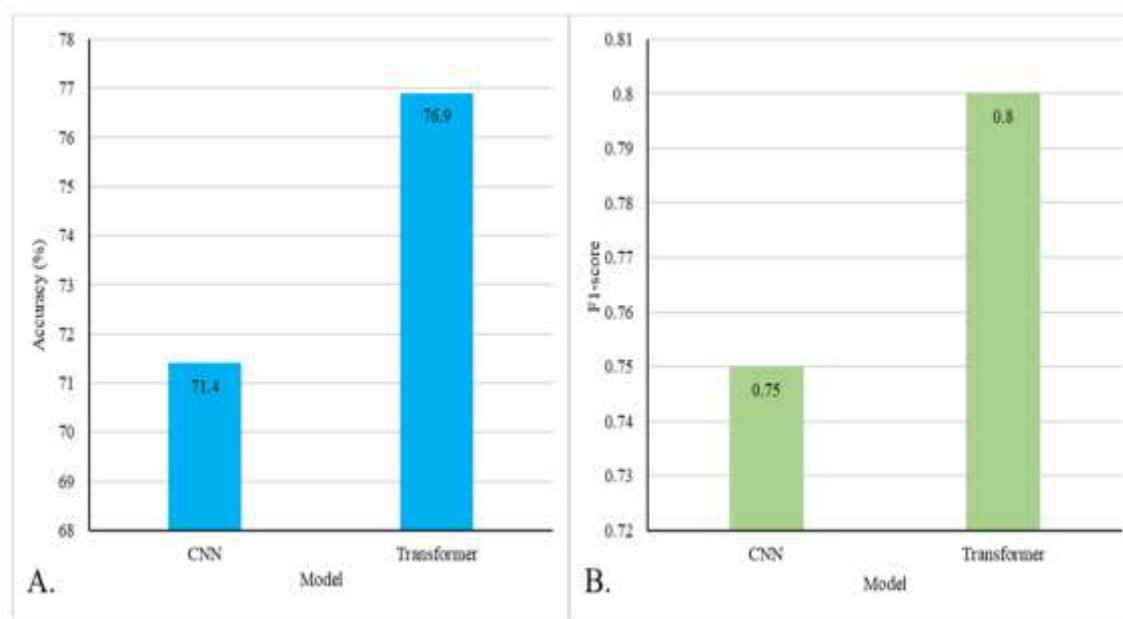


Figure 1. Overall model performance comparison: (A) Accuracy comparison between CNN and Transformer models; (B) F1-score comparison between CNN and Transformer models.

3.2 Performance Across Functional Signal Categories

In order to study the architectural behavior of a variety of linguistic signal transformations, performance was explored in relation to the chosen functional category that was selected in HateCheck. These classes are considered controlled perturbations like spelling variation, semantic negation and contextual quotation. According to Table 2, CNNs have higher F1-scores for categories that are mostly represented by surface-level lexical patterns, including spelling variation and explicit use of slurs. Conversely, the Transformer does much better with more semantically complicated categories, such as negation and quoted hate. These discrepancies are also presented in Figure 2, in which the Transformer proves more stable in the categories that need contextual interpretation.

Table 2. F1-score by Functional Category

Functional Category	CNN F1	Transformer F1
Spelling Variations	0.82	0.78
Explicit Slur Usage	0.85	0.83

Negation Constructs	0.61	0.76
Quoted / Referenced Hate	0.58	0.74
Counter-Speech	0.55	0.72

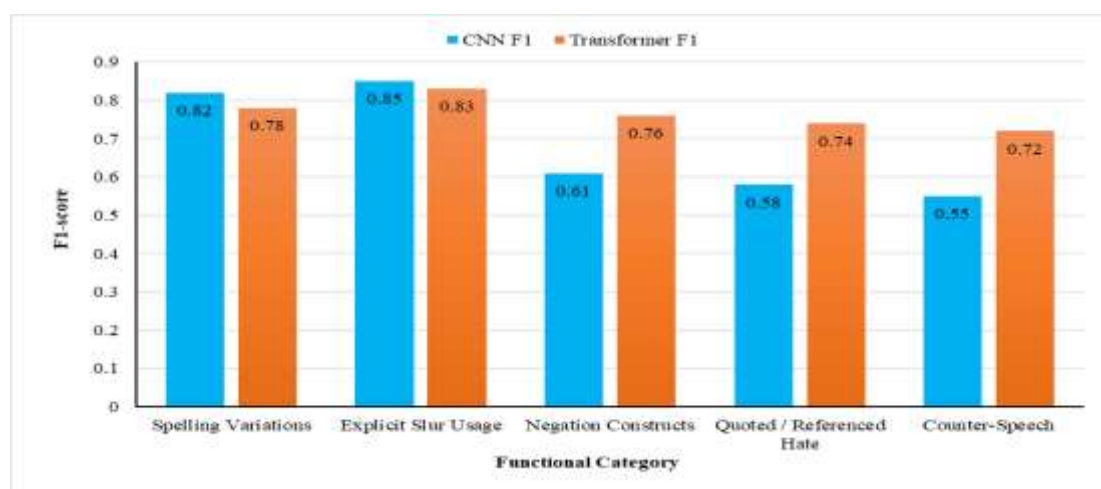


Figure 2. Functional category-wise F1-score comparison.

3.3 Robustness under Spelling-Based Perturbations

To test the sensitivity of the model to character-level noise (adversarial and informal text that are typical on the internet), a tailored robustness analysis was performed. Table 3 indicates that CNNs are more accurate in all sources of noise, and it is more resistant to distortions at character levels. Figure 3 shows the relative performance decrease of the two models at the spelling noise and confirms that the performance of Transformers improves more significantly under the spelling noise, though above a reasonable range.

Table 3. Accuracy under Spelling Noise

Noise Type	CNN Accuracy (%)	Transformer Accuracy (%)
Leetspeak Substitution	84.6	80.2
Character Deletion	81.9	78.4
Space Insertion	83.1	79.6

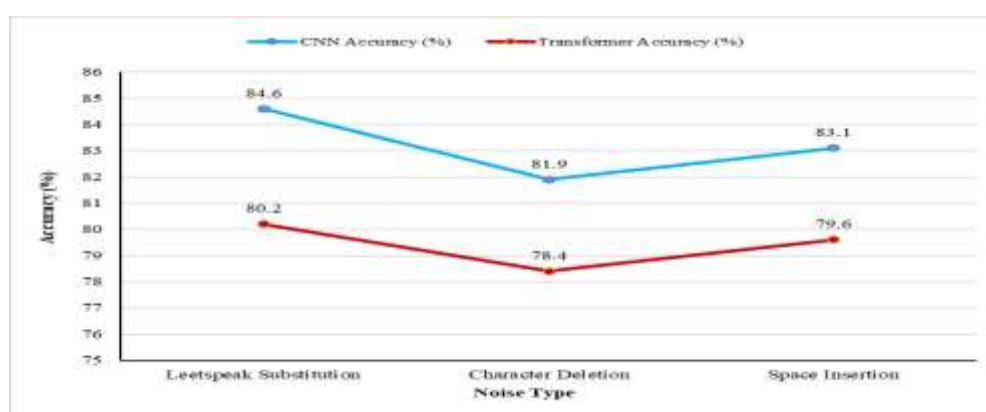


Figure 3. Model robustness under spelling noise perturbations.

3.4 Error Distribution Analysis

In order to get a better insight into the classification behavior, the false positive and false negative rates were analyzed in both architectures. According to Table 4, the False negative rate

of the Transformer is much lower, which minimizes the possibility of not finding hateful content. Although CNNs are somewhat precise, they have a greater false negative, indicating that they are not able to detect semantically carrying hateful utterances. Figure 4 demonstrates that the errors of the Transformer are more balanced than the errors of the CNN.

Table 4. Class-wise Error Rates

Model	False Positive Rate (%)	False Negative Rate (%)
CNN	21.7	28.4
Transformer	18.2	19.6

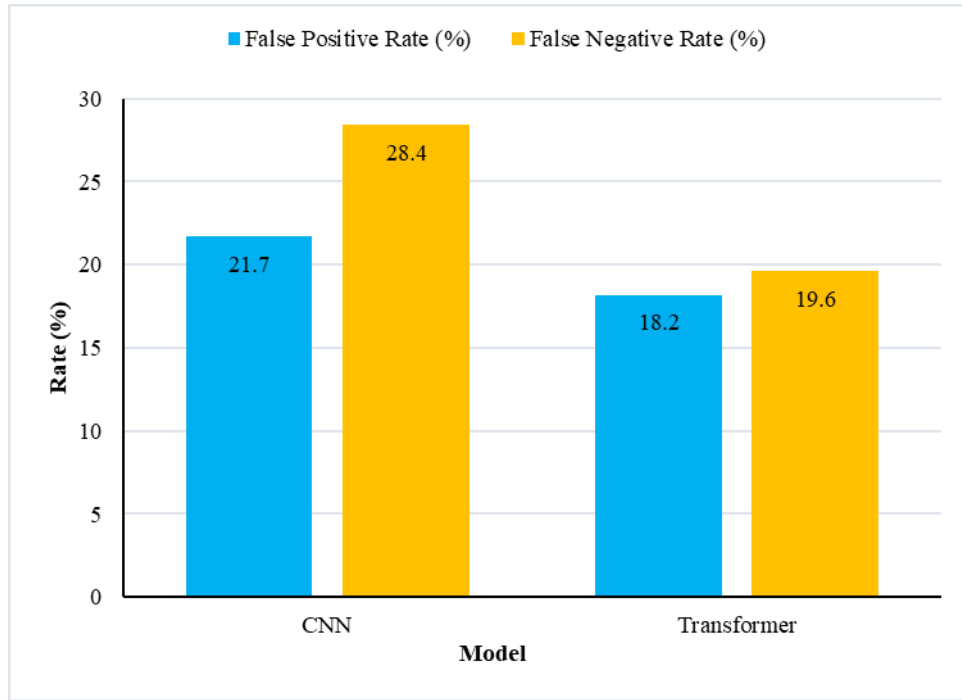


Figure 4. Class-wise error distribution for CNN and Transformer models.

4. Discussion

The findings show that convolutional neural networks and Transformer-based architectures have different and complementary behavior under a controlled linguistic signal perturbation evaluation. On the aggregate level, the Transformer has better accuracy and F1-score, although the false negative rate is significantly lower. It means a greater capability of catching context-dependent and implicitly expressed hateful data, which frequently involves the incorporation of information beyond the local lexical clues. The functionality-based analysis yields a more granular interpretation. Transformer demonstrates regularity in the gains on semantically complex linguistic structures like negation, quoted/quoted hate, and counter-speech. These instances also demand that the model solve the compositional meaning and pragmatic intent, and global dependency modeling is useful in these tasks. Conversely, the CNN has a more decent surface-level perturbation resisting, such as spelling differences and explicit slurs. This is an inductive bias of convolutional architectures to localized pattern detection that is useful when semantic interpretation is not as hard, and discriminative signals are represented in short character sequences. The architectural analysis is further explained by the robustness analysis. The performance of CNN does not deteriorate significantly when the noise goes character-level, but the Transformer is more affected. Spelling perturbations can be viewed as localized distortions, according to which convolutional filters are more robust to salient features than the

attention mechanisms based on tokens. These results support the opinion that robustness is closely correlated with the correspondence of the architectural bias with the type of distortion. The reported benefits of Transformer-based models in semantically challenging signals are in line with the current results of speech and multimodal recognition. The Swin Transformer models of dysarthric speech recognition indicate that attention-based models are suitable to capture long-range relationships in dynamic sequential features in cases where integration of context is necessary [15]. On the same note, the significance of global feature fusion, as a method of decoding subtle affective information, is highlighted in Transformer-based bimodal emotion recognition systems [16]. The complementary nature of convolution and attention is also further evidenced by the evidence of hybrid CNN Transformer models. Analysis of cerebral blood oxygen signals based on CNN-Transformer networks demonstrates that convolutional layers provide robustness to local variations, whereas attention mechanisms enhance sensitivity to global temporal variations [17]. Similar results have been obtained in multimodal physiological emotion recognition, where CNN transformer combinations are superior to single-architecture ones at balancing local and global representation [18].

These findings are applicable to deep learning architectures only. Comparative analysis on recurrent and classical models of modulation recognition also indicates that no one model family is superior to the other in all signal conditions, and error profiles should be measured alongside accuracy [19]. Recent discussion of signal-processing principles in large language models holds that the linguistic inputs can be sensibly studied using terms of noise, invariance, and transformation, and provides a systematic way of studying model behavior [20]. Similar strategies that combine CNNs, Transformers, and graph-based representations to modulate images also indicate that the structural variety of architectures improves the capability of individuals to resist the effects of complicated signal distortions [21]. The current results are consistent with these views in that they empirically show how architectural inductive biases are observed at the controlled linguistic perturbations [22].

The outcomes have a number of implications for designing and implementing NLP systems. To start with, they propose that the choice of architecture be guided by the prevailing nature of the input signal and not the overall benchmark performance in isolation. Transformer-based models can be more useful in applications that involve sensitivity to implicit meaning, contextual framing, or pragmatic intent, and convolutional methods could still be useful in an environment where the surface forms are noisy or adversarial. Second, the results encourage the design of hybrid architectures, which clearly integrate contextual attention with character-level convolution or, at least, safety-critical NLP problems. These designs can offer a principled trade-off between robustness and semantic expressiveness and make fewer trade-offs between the two, as seen with single-architecture systems.

There are a number of limitations that must be admitted. The analysis is performed over a set of diagnostic benchmarks consisting of controlled functional tests, which allow making interpretations, but they might not be as representative of the actual language usage. The research is also limited to fixed architectural setups; the change in tokenization policy, size of a model, or training can affect the performance trends. More work in this direction is required in future to generalize this signal-based evaluation framework to naturalistic and multilingual data, study hybrid CNN-Transformer models, and other strength dimensions like adversarial paraphrasing and domain shift. The inclusion of the analysis of calibration and fairness would also enhance the extrapolability of the results to the deployment of NLP-oriented systems.

5. Conclusion

The complementary properties of attention-based and convolutional architectures are demonstrated under the condition when natural language is regarded as a signal that can be controlled with respect to perturbations. Transformer was more effective overall, as it had better

accuracy and F1-score and a lower false negative rate, which means that it was more sensitive to context-dependent and semantically implicit hate expressions. Convolutional modeling also continued to be beneficial in the case of character-level distortions, and with greater robustness to spelling-based noise, and retains performance when discriminative cues are mostly local. The further functionality analysis revealed that self-attention gains are strongest in constructions that need compositional interpretation, such as negation, quotation and counter-speech, but convolutional feature extraction is competitive in lexicalized and surface-driven ones. The results confirm a view of architecture selection where the error risks and perturbation regimes are the overwhelming factors that dominate the model selection. There are situations where the importance of semantic context and pragmatic framing are essential, and attention-centric systems are desirable, but convolutional components are desirable when the system needs to be resilient to localized input corruption. These findings should be confirmed in future work by using naturalistic and multilingual corpora and analyzing more powerful adversarial transformations, and in hybrid CNN -Transformer layouts, linking the local noise tolerance of local viewpoints with global dependency modelling.

References

1. Liu, Y., Sun, G., Qiu, Y., Zhang, L., Chhatkuli, A. and Van Gool, L., 2021. Transformer in convolutional neural networks. arXiv preprint arXiv:2106.03180, 3.
2. Arkin, E., Yadikar, N., Xu, X., Aysa, A. and Ubul, K., 2023. A survey: object detection methods from CNN to transformer. *Multimedia Tools and Applications*, 82(14), pp.21353-21383.
3. Li, C., Huang, X., Song, R., Qian, R., Liu, X. and Chen, X., 2022. EEG-based seizure prediction via Transformer guided CNN. *Measurement*, 203, p.111948.
4. Thomas, J.B., Chaudhari, S.G. and Verma, N.K., 2023. CNN-based transformer model for fault detection in power system networks. *IEEE Transactions on Instrumentation and Measurement*, 72, pp.1-10.
5. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. and Davison, J., 2020, October. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38-45).
6. Rothman, D. and Gulli, A., 2021. *Transformers for natural language processing* (Vol. 267). Packt Publishing.
7. Tunstall, L., Von Werra, L. and Wolf, T., 2022. *Natural language processing with transformers*. "O'Reilly Media, Inc."
8. Canchila, S., Meneses-Eraso, C., Casanoves-Boix, J., Cortés Pellicer, P. and Castello Sirvent, F., 2024. Natural language processing: An overview of models, transformers and applied practices.
9. Hajri, F. and Fradi, H., 2022, November. Vision transformers for road accident detection from dashboard cameras. In *2022 18th IEEE international conference on advanced video and signal based surveillance (AVSS)* (pp. 1-8). IEEE.
10. Akinpelu, S., Viriri, S. and Haroon Yousaf, M., 2025. Swintser: An improved bilingual speech emotion recognition using shift window transformer. *Cognitive Computation*, 17(4), p.129.
11. Vidyasagar, K.C., Kumar, K.R., Sai, G.A., Ruchita, M. and Saikia, M.J., 2024. Signal to image conversion and convolutional neural networks for physiological signal processing: A review. *Ieee Access*, 12, pp.66726-66764.
12. Moslhi, A.M., Aly, H.H. and ElMessiery, M., 2024. The impact of feature extraction on classification accuracy examined by employing a signal transformer to classify hand gestures using surface electromyography signals. *Sensors*, 24(4), p.1259.

13. Ali, O.S. and Rahim, L., 2025. A Comparative Study of Machine Learning Techniques for AE-Based Corrosion Detection with Emphasis on Transformer Models. *International Journal of Advanced Computer Science & Applications*, 16(7).
14. Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H. and Pierrehumbert, J.B., 2021. HateCheck: Functional tests for hate speech detection models. GitHub repository: HateCheck Dataset. Available at: <https://github.com/paul-rottger/hatecheck-data>.
15. Mahum, R., Ganiyu, I., Hidri, L., El-Sherbeeney, A.M. and Hassan, H., 2025. A novel Swin transformer based framework for speech recognition for dysarthria. *Scientific Reports*, 15(1), p.20070.
16. Soleimani, S. and Al Osman, H., 2025. Transformer-based bimodal emotion recognition: signal measurement and transformer fusion approach. *IEEE Transactions on Instrumentation and Measurement*.
17. Jin, Z., Xing, Z., Wang, Y., Fang, S., Gao, X. and Dong, X., 2023. Research on emotion recognition method of cerebral blood oxygen signal based on CNN-transformer network. *Sensors*, 23(20), p.8643.
18. Sato, M., Huang, Z., Guo, A. and Ma, J., 2025, October. Emotion Recognition based on Multimodal Physiological Signals using a CNN-Transformer Model. In *2025 IEEE Cyber Science and Technology Congress (CyberSciTech)* (pp. 393-400). IEEE.
19. Ansari, S., Mahmoud, S., Majzoub, S., Almajali, E., Jarndal, A. and Bonny, T., 2025. A Novel LSTM Architecture for Automatic Modulation Recognition: Comparative Analysis with Conventional Machine Learning and RNN-Based Approaches. *IEEE Access*.
20. Chen, Q., Zhang, F., Wang, Y., Yu, Q., Lang, G. and Zeng, L., 2025. Bearing fault diagnosis based on efficient cross space multiscale CNN transformer parallelism. *Scientific Reports*, 15(1), p.12344.
21. Verma, P. and Pilanci, M., 2024. Towards signal processing in large language models. *arXiv preprint arXiv:2406.10254*.
22. Wang, D., Lin, M., Zhang, X., Huang, Y. and Zhu, Y., 2023. Automatic modulation classification based on CNN-transformer graph neural network. *Sensors*, 23(16), p.7281.