Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Automated Clinical Text Categorization and Sentimental Analysis

Diya Manoj Poduval
Dept. of CSE
FISAT, Angamaly

Gopika E
Dept. of CSE
FISAT, Angamaly

Maria Manuel
Dept. of CSE
FISAT, Angamaly

Merlin Susan Jacob
Dept. of CSE
FISAT, Angamaly

Ms. Hansa J Thattil
Assistant Professor (Senior Grade)
FISAT, Angamaly

***Abstract*—Healthcare institutions generate vast amounts of unstructured clinical text, making automated analysis essential for efficient decision-making and improved healthcare outcomes. This study proposes a transformer-based approach utilizing Bidirectional Encoder Representations from Transformers (BERT) for both sentiment analysis and medical specialty classification of electronic health records, including clinical notes and discharge summaries. To address class imbalance and improve model performance, the study focuses on the most dominant specialty categories within the dataset. The proposed framework leverages BERT's ability to capture deep contextual and semantic relationships in medical narratives, enabling accurate classification of clinical content as well as effective detection of sentiment polarity. The model is evaluated using standard performance metrics and demonstrates superior results compared to traditional machine learning approaches. The findings highlight the effectiveness of transformer-based models in handling complex medical text and provide a scalable solution for automated clinical text analysis. This approach facilitates faster information extraction, reduces manual workload, and supports enhanced clinical decision-making and patient care.**

***Keywords*—BERT, Sentiment Analysis, Medical Text Classification, Deep Learning, Clinical NLP.**

I. Introduction

The rapid digitization of healthcare systems has led to the widespread adoption of Electronic Health Records (EHRs), generating large volumes of unstructured clinical text. These records contain critical information related to patient conditions, diagnoses, and clinical observations. Extracting meaningful insights from such data is essential for improving clinical decision-making and healthcare quality; however, the complexity of medical language and contextual variability make accurate text classification challenging.

Traditional machine learning techniques often struggle to capture complex contextual relationships, particularly in multi-class classification tasks. Recent advancements in Natural Language Processing (NLP), especially transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT), have significantly improved the ability to understand contextual and semantic patterns in text through bidirectional learning. In this work, a BERT-based framework is proposed for medical specialty classification and sentiment analysis using clinical text data. The dataset, obtained from Kaggle [1], contains multiple specialty classes, from which the 15 most dominant categories were selected to address class imbalance and improve model generalization.

The proposed approach effectively captures contextual features from clinical narratives and demonstrates strong performance in classification tasks. The results highlight the potential of transformer-based models for automated healthcare text analysis, enabling efficient specialty identification and sentiment understanding in large-scale medical records.

II. Literature Study

The growing volume of cancer-related online content demands efficient and accurate methods for emotion analysis and medical text classification. Edara et al. [2] applied traditional machine learning models such as Naïve Bayes, SVM, and Random Forest, achieving reasonable performance but facing limitations in handling large-scale data and contextual interpretation. To overcome these challenges, a distributed Apache Spark framework integrated with LSTM was introduced, incorporating TF–IDF, LDA-based topic modeling, and dimensionality reduction. Experiments on cancer tweets, health news, and biomedical abstracts reported accuracy up to 97.8%, though issues related to noisy social-media data and domain generalization persisted.Monitoring patient–clinician communication is essential, yet manual annotation is time-consuming and inconsistent. Park et al. [3] automated transcript coding using SVM, Naïve Bayes, and Random Forest, with SVM achieving the best accuracy. Although effective, the system struggled with subtle linguistic variations.Eang et al. [4] combined BERT with RNN layers to enhance sentiment classification by capturing both contextual

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

semantics and temporal dependencies. The hybrid model outperformed baseline methods on the SST-2 dataset but required higher computational resources. Mollaei et al. [5] reviewed ML-based NLP techniques for biomedical text mining, highlighting the shift from traditional models to deep learning approaches such as CNNs, BiLSTM-CRF, and BioBERT. The study emphasized explainable AI and domain-adaptive models for reliable clinical applications.Yang and Emmert-Streib [6] proposed a threshold-learning CNN for multi-label EHR classification, improving prediction accuracy on the MIMIC-III dataset. While effective in handling label imbalance, rare conditions remained challenging.

Carcone et al. [7] automated behavioral coding in clinician–patient communication using models such as SVM, CNN, Random Forest, and Naïve Bayes. SVM achieved the best performance, particularly with enriched lexical and contextual features, and demonstrated generalization across datasets. However, detecting subtle communication behaviors remained challenging. Waheeb et al. [8] applied sentiment analysis to discharge summaries for healthcare quality evaluation. Using models including SVM, Random Forest, CNN, and LSTM with TF–IDF and Word2Vec features, deep learning approaches—especially CNN and LSTM—achieved superior performance. Nonetheless, limited data and domain-specific terminology posed challenges.Wang et al. [9] introduced a weak supervision framework for clinical text classification, generating noisy labels through rule-based heuristics and combining them with deep word embeddings. CNN achieved the highest F1 scores, though performance declined for complex categories. Qing et al. [10] proposed a hybrid CNN–BiGRU model with dual-attention mechanisms for medical text classification. The architecture outperformed traditional and earlier deep learning methods across multiple datasets, offering improved interpretability despite higher computational cost.

III. Methodology

A. Overview

The proposed framework consists of two major components: (1) Medical Specialty Classification and (2) Sentiment Analysis of clinical transcriptions. Both tasks were implemented on the same preprocessed medical transcription dataset to ensure consistency and fair comparison across models. The overall system integrates preprocessing, feature extraction, model training, and performance evaluation stages.

B. Dataset and Preprocessing

The medical transcription dataset was preprocessed to ensure data quality and consistency. Null entries and incomplete records were removed, and only the 15 medical specialties were selected to address class imbalance and maintain uniformity across experiments. This selection improved model generalization and ensured balanced representation.

To standardize the clinical text, preprocessing steps such as lowercasing, removal of extra whitespaces and formatting characters, and encoding of categorical labels were applied. These steps ensured a consistent textual representation for effective model training.

To prevent target leakage, a keyword masking strategy was introduced. Clinical transcripts often contain explicit mentions of specialty names that directly reveal the target class. Such keywords were identified, cleaned, and replaced with a neutral token `[MASK]` using regular expression-based matching. This approach ensures that models learn contextual patterns rather than relying on direct keyword associations. For feature extraction, traditional machine learning models used TF-IDF vectorization to convert text into numerical representations. The TF-IDF score is defined as:

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t) \tag{1}$$

For deep learning and transformer-based models such as CNN, LSTM, BERT, and BioClinicalBERT, tokenization was performed using model-specific tokenizers. Input sequences were padded to a maximum length of 512 tokens, and attention masks were used to differentiate valid tokens from padding.

C. System Architecture

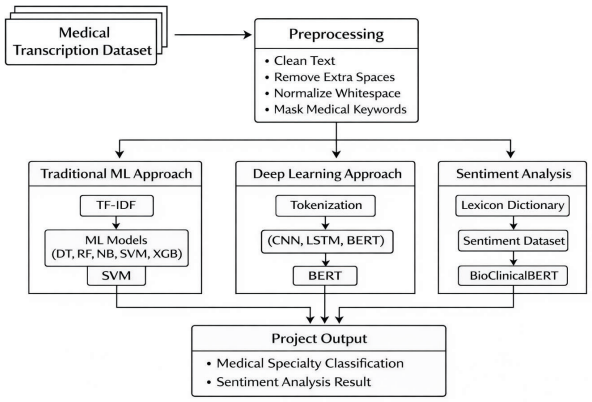

Fig. 1. System Architecture for Specialty Classification and Sentiment Analysis

Fig. 1 illustrates the overall architecture of the proposed system for clinical text analysis. The process begins with a medical transcription dataset, which undergoes preprocessing steps such as text cleaning, whitespace normalization, and keyword masking. The processed data is then fed into three parallel modules: a traditional machine learning approach using TF-IDF and classifiers like SVM, a deep learning approach leveraging tokenization and models such as CNN, LSTM, and BERT, and a sentiment analysis module based on lexicon methods and BioClinicalBERT. The outputs from these modules are combined to produce final results, including medical specialty classification and sentiment analysis outcomes.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

D. Medical Specialty Classification Models

A range of classical machine learning classifiers were implemented for medical specialty classification. Decision Tree was used as a baseline model due to its interpretability, though it is prone to overfitting. Random Forest improved robustness by combining multiple decision trees, while Naïve Bayes provided efficient performance on high-dimensional TF-IDF features. Support Vector Machine (SVM) was employed for its effectiveness in handling high-dimensional text data by maximizing class separation. Additionally, XGBoost was used for its ability to enhance performance through iterative learning. All machine learning models utilized TF-IDF features as input representations.

In addition to classical approaches, deep learning models were implemented to capture semantic and contextual relationships in clinical text. CNN was used for extracting local features, while LSTM modeled sequential dependencies and long-range context. A transformer-based model, BERT, was employed for its bidirectional contextual understanding and was fine-tuned for multi-class medical specialty classification, resulting in improved performance. For BERT, tokenization was performed with a maximum sequence length of 512 tokens. The [CLS] token embedding was used as the aggregate representation of the input and passed to a fully connected layer for final classification.

E. Sentiment Analysis Framework

In addition to medical specialty classification, sentiment analysis was performed to determine the emotional polarity present in clinical text. Since the dataset did not contain predefined sentiment labels, a lexicon-based approach was adopted to generate initial annotations. The VADER sentiment analyzer was used to compute a compound polarity score for each transcription, based on which texts were classified into Positive (score ≥ 0.05), Negative (score ≤ -0.05), and Neutral categories. These generated labels were stored as the target variable for subsequent model training.

For sentiment classification, BioClinicalBERT was utilized due to its domain-specific pretraining on clinical datasets such as MIMIC-III and PubMed. This enables the model to effectively understand medical terminology, abbreviations, and contextual relationships within clinical narratives. The model processes tokenized input using WordPiece embeddings, and contextual representations are generated through multiple transformer layers, followed by a classification layer, as illustrated in Fig. 1.

The dataset was divided into training and testing sets using stratified sampling to preserve class distribution. The model was fine-tuned using cross-entropy loss and optimized with the Adam optimizer with weight decay for stable convergence. Hyperparameters such as learning rate and batch size were selected empirically. The performance of the model was evaluated using accuracy and weighted F1-score to ensure reliable sentiment classification.

IV. Results and Discussion

A. Medical Specialty Classification Results

The performance of traditional machine learning models and deep learning architectures was evaluated using Accuracy, Precision, Recall, and F1-score. Table I presents the comparative performance of all implemented models.

TABLE I
Performance Comparison of Classification Models

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	43	47.46	42.98	43.78
Random Forest	60.17	62.38	60.17	60.34
Naive Bayes	74	79	74	70
LSTM	82	81.77	81.87	81.57
BERT	86	87	86	86
XG boost	75	77	75	74
SVM	89	89	89	88
CNN	71	69	71	67

From the results, it is evident that classical tree-based models such as Decision Tree achieved relatively low performance due to limitations in handling high-dimensional sparse textual features. Random Forest improved generalization through ensemble learning but remained limited by feature representation constraints.

Naïve Bayes and XGBoost demonstrated moderate performance, benefiting from probabilistic modeling and boosting strategies. However, these models rely heavily on handcrafted TF-IDF features and lack contextual understanding.

Deep learning models showed significant improvements. The CNN model captured local textual patterns effectively, achieving competitive performance. The LSTM model further improved accuracy by modeling sequential dependencies in medical transcripts, highlighting the importance of contextual information in clinical narratives.

Transformer-based BERT achieved strong performance with 86% accuracy, demonstrating the advantage of contextual embeddings over traditional feature extraction techniques. Interestingly, the SVM classifier achieved the highest accuracy of 89%, indicating that for moderately sized datasets with well-optimized TF-IDF features, classical margin-based classifiers can outperform deep transformer architectures.

Overall, the comparative analysis confirms that contextual learning improves performance, but optimized classical models remain competitive depending on dataset size and feature engineering quality.

The confusion matrices for SVM and BERT are illustrated in Fig. 2 and Fig. 3. The SVM classifier demonstrates strong diagonal dominance, indicating accurate class predictions across most specialties. BERT also shows high true positive rates; however, minor inter-class confusion is observed in closely related specialties, highlighting the complexity of medical terminology.

The ROC curves for SVM and BERT are illustrated in Fig. 4 and Fig. 5. Since the task involves multi-class classification across 15 specialties, the One-vs-Rest (OvR) strategy was

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

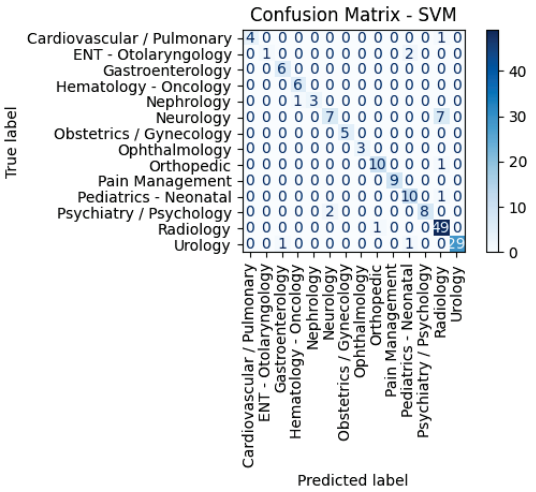

Fig. 2. Confusion matrix for SVM-based medical specialty classification.

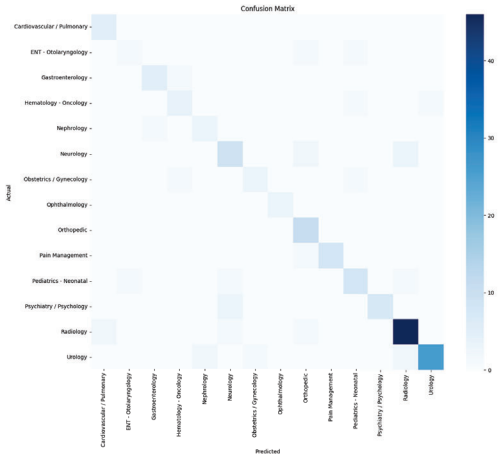

Fig. 3. Confusion matrix for BERT-based medical specialty classification.

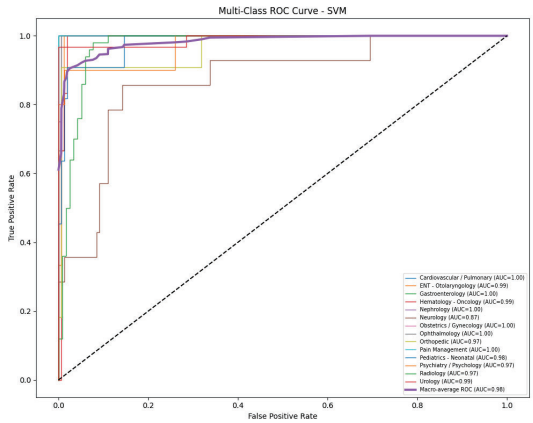

Fig. 4. ROC curve for SVM model.

employed to compute class-wise ROC curves, and macro-averaged AUC was used for overall comparison.

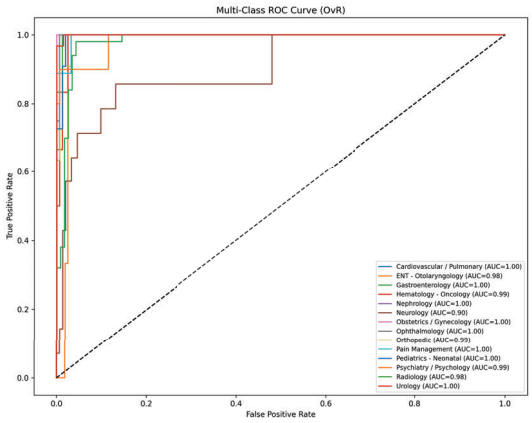

Fig. 5. ROC curve for SVM model.

The SVM model demonstrates a consistently higher Area Under the Curve (AUC), indicating strong discriminative capability using TF-IDF features. BERT also achieves competitive AUC performance by leveraging contextual embeddings; however, the relatively moderate dataset size may limit the full advantage of transformer-based representations.

Overall, the ROC analysis confirms the robustness of both models, while highlighting the superior margin-based separation achieved by SVM in this experimental setting.

B. Sentiment Analysis Results

In addition to medical specialty classification, sentiment analysis was performed using transformer-based models. Since the dataset did not contain predefined sentiment labels, initial labels (Negative, Neutral, and Positive) were generated using the VADER sentiment analyzer and used for model training. Three transformer-based models, namely BioClinicalBERT, BioLinkBERT, and DeBERTa, were evaluated for the sentiment classification task.

The performance comparison of these models is presented in Table II. Among the evaluated models, BioClinicalBERT achieved the highest accuracy and F1-score, indicating its superior ability to capture contextual and domain-specific features in clinical text. Based on this performance, BioClinicalBERT was selected as the final model for sentiment analysis in the proposed system.

TABLE II
PERFORMANCE COMPARISON OF SENTIMENT MODELS

Model	Accuracy	F1 Score
BioClinicalBERT	0.8113	0.8069
BioLinkBERT	0.8008	0.7969
DeBERTa	0.5241	0.3605

1) Overall Sentiment Distribution: The overall sentiment distribution across the selected 15 medical specialties is illustrated in Fig. 6. The results show that most clinical reports fall under the Negative and Positive sentiment categories, while Neutral instances are comparatively fewer.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

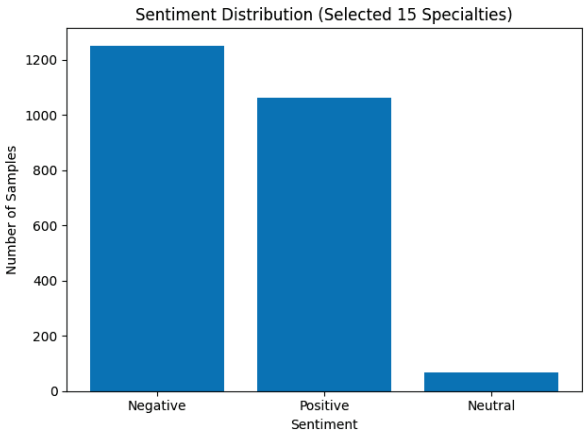

Fig. 6. Overall sentiment distribution across the selected 15 medical specialties.

2) Sentiment Distribution Across Medical Specialties: The sentiment distribution across different medical specialties is presented in Fig. 7. It is observed that certain specialties exhibit a higher proportion of negative sentiment, reflecting the nature of clinical documentation.

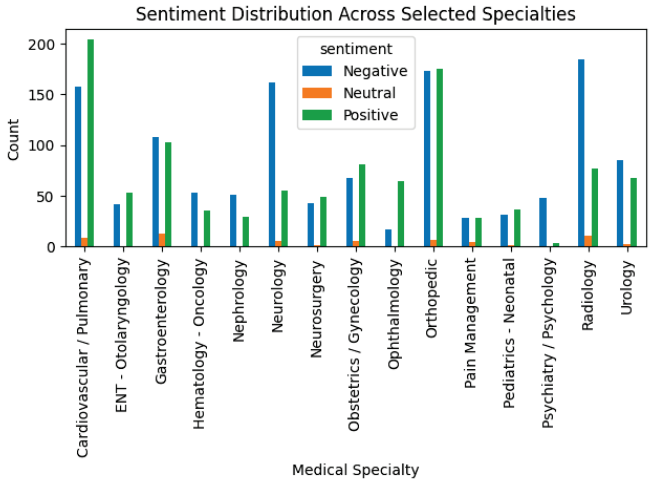

Fig. 7. Sentiment distribution across the selected medical specialties.

C. Comparative Discussion

The experimental results highlight that transformer-based models such as BERT and BioClinicalBERT significantly improve contextual understanding compared to traditional approaches when applied to clinical text. Support Vector Machine (SVM) demonstrated strong and consistent performance, confirming that TF-IDF–based feature engineering remains highly effective for medical specialty classification. While SVM provided an efficient and reliable baseline with lower computational complexity, domain-specific transformers like BioClinicalBERT were better at capturing semantic nuances and subtle contextual cues, particularly in sentiment analysis. However, sentiment classification in clinical documentation remains challenging due to the predominance of neutral language. Overall, the findings indicate that both traditional and transformer-based models are valuable, with their effectiveness depending on dataset characteristics, task complexity, and computational constraints.

V. REFERENCES

[1] https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions
[2] D. C. Edara, L. P. Vanukuri, V. Sistla, and V. K. K. Kolli, "Sentiment Analysis and Text Categorization of Cancer Medical Records with LSTM," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, pp. 5309–5325, 2023.
[3] G. Park, J. T. Rayz, and C. G. Shields, "Towards the Automatic Coding of Medical Transcripts to Improve Patient-Centered Communication," in *Proceedings of the IEEE International Conference on Healthcare Informatics (ICHI)*, Purdue University, 2019.
[4] C. Eang and S. Lee, "Improving the Accuracy and Effectiveness of Text Classification Based on the Integration of the BERT Model and a Recurrent Neural Network (RNN_BERT_Based)," *Applied Sciences*, vol. 14, no. 18, p. 8388, Sep. 2024, doi: 10.3390/app14188388.
[5] N. Mollaei, C. Cepeda, J. Rodrigues, and H. Gamboa, "Biomedical Text Mining: Applicability of Machine Learning-based Natural Language Processing in Medical Database," in *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2022) - Vol. 4: BIOSIGNALS*, pp. 159–166, 2022.
[6] Z. Yang and F. Emmert-Streib, "Threshold-Learned CNN for Multi-Label Text Classification of Electronic Health Records," *IEEE Access*, vol. 11, pp. 93402–93418, 2023, doi: 10.1109/ACCESS.2023.3309157.
[7] A. I. Carcone, M. Hasan, G. L. Alexander, M. Dong, S. Eggly, K. B. Hartlieb, S. Naar, K. MacDonell, and A. Kotov, "Developing Machine Learning Models for Behavioral Coding," *Journal of Pediatric Psychology*, vol. 44, no. 3, pp. 289–299, 2019.
[8] S. A. Waheeb, N. A. Khan, B. Chen, and X. Shang, "Machine Learning Based Sentiment Text Classification for Evaluating Treatment Quality of Discharge Summary," *Applied Sciences*, vol. 10, no. 10, May 2020.
[9] Y. Wang, S. Sohn, S. Liu, F. Shen, L. Wang, E. J. Atkinson, S. Amin, and H. Liu, "A Clinical Text Classification Paradigm Using Weak Supervision and Deep Representation," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, pp. 1–13, 2019.
[10] L. Qing, W. Linhong, and D. Xuehai, "A Novel Neural Network-Based Method for Medical Text Classification," *Future Internet*, vol. 11, no. 12, pp. 1–13, 2019.
[11] M. Khadhraoui, H. Bellaaj, M. B. Ammar, H. Hamam, and M. Jmaiel, "Survey of BERT-base models for scientific text classification: COVID-19 case study," *Applied Sciences*, vol. 12, no. 6, p. 2891, 2022.
[12] A. Moharram, S. Altamimi, and R. Alshammari, "Data analytics and predictive modeling for appointments no-show at a tertiary care hospital," in *Proc. 2021 1st Int. Conf. Artificial Intelligence and Data Analytics (CAIDA)*, Riyadh, Saudi Arabia, 2021, pp. 275–277.
[13] C. Palmirotta, S. Aresta, P. Battista, S. Tagliente, G. Lagravinese, D. Mongelli, and C. Salvatore, "Unveiling the diagnostic potential of linguistic markers in identifying individuals with Parkinson's disease through artificial intelligence: A systematic review," *Brain Sciences*, vol. 14, no. 2, Art. no. 137, 2024.
[14] A. Feder, D. Vainstein, R. Rosenfeld, T. Hartman, A. Hassidim, and Y. Matias, "Active deep learning to detect demographic traits in free-form clinical notes," *Journal of Biomedical Informatics*, vol. 107, Art. no. 103436, 2020.
[15] W. H. Bangyal, R. Qasim, N. U. Rehman, Z. Ahmad, H. Dar, L. Rukhsar, and J. Ahmad, "Detection of fake news text classification on COVID-19 using deep learning approaches," *Computational and Mathematical Methods in Medicine*, vol. 2021, Art. no. 5514220, 2021.
[16] P. Guleria, "NLP-based clinical text classification and sentiment analyses of complex medical transcripts using transformer model and machine learning classifiers," *Neural Computing and Applications*, vol. 37, pp. 341–366, 2025, doi: 10.1007/s00521-024-10482-x.