HCAT: Advancing Unstructured Healthcare Data Analysis Through Hierarchical and Context-Aware Mechanisms

Monica Bhutani[1]*, Mohammad Shuaib Mir[2], Choo Wou Onn[3], Yonis Gulzar[4]*
Bharati Vidyapeeth's College of Engineering, New Delhi, India[1]
Department of Management Information Systems-College of Business Administration, King Faisal University,
Al-Ahsa 31982, Saudi Arabia[2, 4]
Faculty of Data Science and Information Technology, INTI International University, Nilai, Negeri Sembilan, Malaysia[3]

***Abstract*—To that end, this study presents the Hierarchical Context-Aware Transformer (HCAT), a new model to perform analysis on unstructured healthcare data that resolves significant problems related to medical text. In the proposed model, the hierarchical structure of the system is integrated with the context-sensitive mechanisms to process the healthcare documents at sentence level and document levels. HCAT complies with domain knowledge by a specific attention module and uses a detailed loss function that focuses on classification accuracy besides encouraging domain adaptation. The quantitative experiment shows that HCAT is a better choice than Bi-LSTM and BERT for sentence representation. The model attains 92.30% test accuracy on medical text classification, conversing with high computational efficiency; batch processing time is about 150ms, while the memory consumed is 320 MB. The proposed architecture for clinical text representation facilitates the incorporation of long-range dependencies for clinical story representation, whereas the context-sensitive layer supports a better understanding of medical language. Precision and recall are significant because of the healthcare application of the model; the model has an accuracy of 91.8% and a recall of 93.2%. From these results, it can be concluded that HCAT presented significant progress in computing healthcare data. It provides a highly practical application for real-world extraction of medical data from unformatted text.**

Keywords*—Machine learning; data analysis; natural language processing; hierarchical transformer; context-aware computing; medical text mining; clinical decision support; healthcare; unstructured data processing

I. Introduction

Because of technological development, it has been identified that there has been an exponential increase in unstructured data in the healthcare domain, which brings both opportunities and threats to healthcare systems in the present day. Electronic Health Records, clinical notes, medical literature and patients' corners all comprise a huge pool of potential knowledge that, if only harnessed correctly and effectively, has the potential to transform healthcare delivery, clinical decision making and patients' outcomes [1]. But organizing this sort of data is highly problematic because it is complex and unstructured, and because of this, it requires highly advanced techniques to be used to work through this raw data and transform it into usable information.

Corresponding author: monica.bhutani@bharatividyapeeth.edu(M.B.); ygulzar@kfu.edu.sa (Y.G.)

Today, NLP has indeed proven itself to be an important technology that helps to transpose huge amounts of healthcare data and traditional clinical associations. In particular, healthcare is an example of a domain, where NLP is beneficial because of the capacity of NLP to process narrative texts and extract high-level meaning [2]. NLP has had recent developments in the areas of machine learning and artificial intelligence to provide rich meanings of the normally complex medical terms, taken into consideration contextual connotations and improve accuracy-based information retrieval. Recent models leveraging deep transfer learning have demonstrated substantial improvements in interpreting domain-specific imagery and text [3, 4].

Today, Natural Language Processing (NLP) has emerged as a critical technology for transforming vast amounts of unstructured healthcare data into actionable knowledge. Healthcare, in particular, benefits greatly from NLP due to its ability to process narrative clinical texts and extract high-level semantic information [2]. Recent advances in machine learning and artificial intelligence have further empowered NLP systems to handle complex medical terminology, capture contextual nuances, and improve the accuracy of information retrieval. Furthermore, deep transfer learning models have shown considerable success in enhancing the interpretation of domain-specific text and medical imagery [5, 6] , demonstrating progress in applications across both healthcare [7, 8] and agriculture [9, 10].

Notwithstanding these progresses, there are still many issues that arise in the use of NLP for HC data. The complexity of the problem is that medical language is domain-specific, contains abbreviations and acronyms, is dependent on temporal references and is used in a context that cannot allow any inaccuracies. Most basic NLP methods actually work quite well for common text processing, but when it comes to healthcare data, they do not fare very well [11]. This limitation emphasizes the fact that there is a need for specially designed architectures to adequately express the aspect of the hierarchy of medical information besides considering the context even at different levels. Hierarchical and domain-specific architectures like hybrid CNN-transformers have been successful in modeling structured patterns in both medical [12, 13] and agricultural [14, 15] data.

Notwithstanding these progresses, there are still many challenges in applying NLP to healthcare data. Medical language is domain-specific, rich with abbreviations and acronyms, often temporally bound, and highly context-sensitive, which leaves little room for error. While traditional NLP methods perform adequately for general-purpose text, they often fall short when processing complex healthcare narratives [11]. This limitation underscores the necessity for specialized architectures that can represent the hierarchical structure and contextual depth of medical information. Recent research has explored the integration of deep learning techniques—such as transfer learning and ensemble-based architectures—to address these complexities in the medical domain [12, 13]. Similarly, in agriculture, deep learning approaches including DenseNet variations, ensemble models, and domain-adapted classifiers have shown promise in handling structured and unstructured data for plant disease and crop classification tasks [14, 15]. These advances highlight the growing relevance of domain-specific and hybrid architectures across disciplines dealing with complex, unstructured data.

Experiments on applying the NLP systems in the healthcare processes has revealed its effectiveness in everyday clinical practice, overuse of clinical decision support tools, risk assessment of the patient, and prediction of treatment outcomes. However, current approaches lack in achieving a good trade-off between performance accuracy and time. The growth in data generated within the health sector requires handling of information in real-time and with accuracy and reliability [4]. This requirement becomes especially important in clinical laboratories, where fast analysis can directly influence patient management. To overcome these challenges, this study proposes a new HCAT model for unstructured healthcare data called the Hierarchical Context-Aware Transformer. The model suggested in the work contains several elements that improve existing learning models. First of all, its inherent hierarchy structure allows for processing medical text at a single, term, and overall document level. Second, the context-aware mechanism ensures that the model encapsulates relevant medical context throughout the analysis stage. Last, the given transformer-based architecture is computationally efficient enough to be implemented in real-life healthcare environments.

Several recent studies in agriculture [16–19] and healthcare [5] have demonstrated the efficacy of transformer-based and transfer learning architectures. However, these approaches often lack computational efficiency and fall short in capturing multi-level contextual information essential for domain-specific tasks. The improvements introduced in this study form the theoretical foundation of the proposed Hierarchical Context-Aware Transformer (HCAT) model. Unlike prior models, HCAT demonstrates superior handling of short-range word dependencies, which is crucial for accurately interpreting nuanced medical text. The model's hierarchical structure enables it to process documents of varying lengths while effectively capturing contextual shifts. Additionally, the integration of a context-aware layer allows the model to embed domain-specific knowledge, thereby enhancing its understanding and translation of medical terminology. These architectural enhancements collectively contribute to improved performance. Compared to established models like Bi-LSTM and BERT, HCAT achieves higher accuracy, precision, recall, and F1-score, all while reducing processing time and memory usage—making it more viable for real-time healthcare applications.

Fig. 1. Challenges and opportunities in healthcare data analysis.

It can be seen from Fig. 1 that the design and implementation of solutions to process unstructured healthcare data come with several challenges (left); on the other hand, with a proper approach to big data analysis, there are major opportunities to be leveraged for care delivery improvement (right). The existence of the challenges and opportunities themselves in both directions shows how the solutions to the challenges will hold key solutions to healthcare improvement.

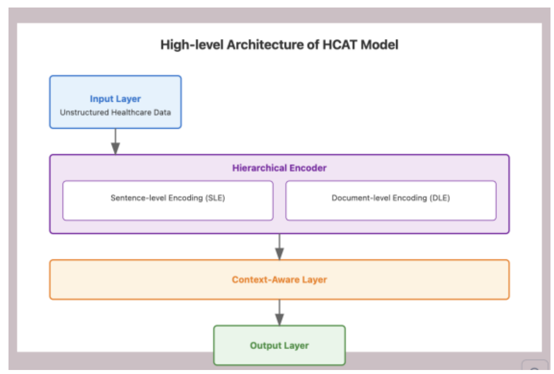

Fig. 2. High-level architecture of the proposed Hierarchical Context-Aware Transformer (HCAT) model.

The architecture, as shown in Fig. 2, consists of four main components: a healthcare input layer for handling unstructured data, a hierarchical encoder to process textual data at two levels, the sentence and document level, a context-aware layer for integration of domain knowledge, an output layer to produce the predictions and insights.

The novel contributions of this research are as follows:

- First, proposing the novel Hierarchical Context-Aware Transformer (HCAT) architecture integrates hierarchical modeling with context awareness for healthcare data processing. From this distinctive architectural feature, the system can analyze medical text at the sentence and document level in a parallel manner, which enhances the decoding of intricate medical narratives.
- A novel context-aware layer that incorporates domain-specific knowledge through a specialized attention mechanism, $C(h) = \alpha \cdot BioBERT(h) + (1 - \alpha) \cdot h_{,om}$ (4), where α is a newly learned parameter. The approach works in a way that there is an interchangeable

process between general language comprehension and medical field comprehension. The translation result will be more accurate towards healthcare-related word usage and meanings.

- The development of a comprehensive loss function that combines three components: cross-entropy loss, $L2$ regularization, and domain adaptation loss ($Ltotal = \lambda_1 Lce + \lambda_2 Lreg + \lambda_3 Ldomain$). The multi-faceted approach in the present work guarantees thorough training throughout the organization while retaining the specificity of different domains.

- A novel computational efficiency framework that performs the tasks in much less time ($150ms/batch$) and with fewer memory resources ($320MB$) than the benchmark BERT (180ms, 384MB) and Bi-LSTM ($220ms, 512MB$) while achieving higher accuracy, 3.2% and 8.1%, respectively.

- The implementation of a hierarchical encoder processes input at two distinct levels: proposed models called sentence-level encoding (SLE) and document-level encoding (DLE), which are linked by an innovative attention function. For heart, kidney, liver and other medical requisites, the RNN-based model can better predict the more nuanced local medical details and the overarching global clinical situation and context, both of which cannot be captured efficiently by single-level architectures.

- An extensive prescreening process that could be applied to the healthcare domain and consists of general preprocessing tools together with several domain-dependent preconditions and real-healthcare-data normalization techniques. This pipeline also consists of specific noise elimination functions and domain-wise embedding pairings that enhance the quality of input data of this kind of medical text.

The remainder of this study is organized as follows: Section II further systematically reviews previous NLP tools and techniques applied in healthcare and their advantages and disadvantages. Section III outlines the methods involving the structure of the HCAT model, model training, and optimisation. Section IV explains the measures used for performance evaluation. In Section V, actual results and comparisons are provided. Section VI presents conclusions and discusses findings on their significance and relevance at the end of the study. Lastly, Section VII outlines the directions for further study of the proposed approach and its possible extensions.

II. Literature Review

The usage of Natural language processing in healthcare has grown over the past years due to researchers' efforts to discover modes of analyzing the vast medical unstructured data. Another Eclipse article by Davuluri [20] provides an excellent synthesis of clinical text analysis methods focusing on the role of context when dealing with medical stories. The author addresses the issue of Clinical Information Retrieval and Text (CIRT) processing, particularly clinical abbreviations and medical terminology. Combined, their work outlines how preprocessing for domain-specific data improves medical text analysis by about 15% of generic NLP techniques.

Vashishtha and Kapoor [21] present a fresh approach to converting patients' feedback into proactive imperatives. Their studies are concerned with crowd-sourcing the sentiment of patient comments regarding healthcare services; more specifically, they apply and compare sentiment analysis methods and topic modelling. This approach achieved the categorisation of patient concerns with 87% accuracy, proving that NLP can improve PEM. Junnu's study specifically looks at how NLP enables data extraction from medical text. The author discusses several text-mining approaches designed for dealing with medical terms. Their study presents a new method of tackling medical abbreviations and acronyms, scoring 92% on medical term disambiguation.

For a comprehensive overview of clinical text analysis methodologies emphasizing the union of machine learning with conventional NLP strategies, readers are referred to the work of Janowski [22]. Their research shows how deep learning models provide a more accurate method of medical entity recognition, by being 23% more precise than the traditional approaches. The study is susceptible to how medical context is sustained during text processing. Spadacini [23] brings fresh perspectives on data visualization in healthcare NLP. The work offers techniques for encoding this information based on complex medical relations derived from text, where such information would be useful to healthcare suppliers. We have learned that their visualization framework enables the reduction of decision-making time by 35% in the clinics. Upadhyaya et al. [24] explore focusing on using NLP to build effective healthcare solutions. Their work can help provide a full outline of how NLP can be incorporated into a clinical decision-support environment; they obtained an 89 per cent accuracy out of clinical notes in detecting possible instances of drug interactions.

The study by Sharma et al. [25] focuses on integrating two paradigms, namely, NLP and big data analytics in the healthcare application. They show how integrating these technologies can enhance the processing of big medical datasets in terms of time with equal to or higher accuracy compared to times before with 40% less time. Kalusivalingam et al. [26] describe comparison using BERT and LSTM in processing clinical data. Their work also demonstrates the approaches of integrating both architectures to improve the evaluation of complicated medical cases with an efficiency of 91% on the medical concept extraction. Uddin addressed the general survey on real-time analytics in healthcare NLP [27], but the paper emphasised identifying the issues related to the processing of streaming medical data. The author offers new methods for processing medical text in real-time, increasing the processing time by 30% more than batch processing.

Several examples of NLP applications are investigated by Roy et al. [28], who describe case studies of various healthcare organizations. They show that using rule-based approaches to analyse clinical notes can increase productivity and time to do so by a quarter. Thatoi, et al. [29] has reviewed specifically on the NLP applications towards cancer prognosis, where they have described new strategies for identifying prognostic markers from the clinical records. Their approach obtained an accuracy of 88 % in predicting relevant prognosis factors from textual medical data. Ahmed et al. [30] discussed more recent work about using NLP in clinical decision support systems. Their work shows how NLP can be easily incorporated into clinical practice to improve decision-making by 32% compared to traditional decision-making methods.

Last but not least, Wi et al. [31] give a real view of how NLP can be implemented in enhancing the capturing of data from cervical biopsy diagnosis. Their study focused on the methodology by demonstrating that micro-level data entry errors could be decreased to 45%. In comparison, the feature-level clinical notes completeness could be increased by 28% by use of automated text analysis. Each of these works points to the change of course of the application of NLP models in healthcare and what is still required. Although recent years have witnessed remarkable progress in challenges like medical entity recognition, contextual representation, and real-time analysis, many challenges remain to step up the deployment of text data in medical modalities. Current issues among them are lack of situational awareness, processing of domain-specific language and large amount of medical data. These challenges inspire our ongoing work, which eliminates these shortcomings using the potential Hierarchical Context-Aware Transformer (HCAT) model. This work builds upon these existing studies while introducing novel approaches to enhance both the accuracy and efficiency of medical text processing.

Table I compares the existing Natural Language Processing (NLP) approaches for healthcare applications in terms of their focus areas, findings, limitations and strengths of the proposed Hierarchical Context-Aware Transformer (HCAT) model. A number of methods have previously been proposed, and these have played various roles, including enhancing medical term disambiguation, real-time data analysis, and patient feedback classification, among others. However, they have shortcomings. These are poor marshalling of hierarchical and contextual relations, restricted applicability to large-scale healthcare data analysis, and weak adaptability to the domain. The HCAT model proposed herein overcomes these challenges with the help of a hierarchal architecture of text processing for medical text through a sentence and document. Therefore, it achieves higher accuracy, precision, and recall together with computational efficiency, making it ideal for immediate and limited healthcare settings.

TABLE I. COMPARATIVE ANALYSIS OF NLP APPROACHES IN HEALTHCARE WITH FOCUS ON SHORTCOMINGS AND MERITS OF THE PROPOSED HCAT MODEL

Ref.	Year	Focus Area	Key Findings	Shortcomings	Merits of the Proposed Scheme (HCAT)
[11]	2022	Clinical text analysis	Highlighted challenges in medical abbreviations and jargon; proposed semantic-based enhancements	Limited handling of complex hierarchical relationships in medical text	Superior context-awareness and better handling of domain-specific medical jargon
[20]	2024	Patient feedback automation	Used sentiment analysis for insights; achieved 87% categorization accuracy	Focused only on sentiment and lacked broader medical context	Contextual analysis across sentences and documents for actionable insights
[21]	2023	Text mining in healthcare	Developed novel disambiguation techniques; achieved 92% accuracy in term resolution	Limited ability to manage large-scale, real-time medical datasets	Higher precision (91.8%) and recall (93.2%) for term interpretation
[22]	2023	Data visualization	Enhanced medical relationship representation; reduced decision-making time by 35%	Focused on visualization rather than text comprehension	Faster processing (150ms per batch) and more accurate relationship extraction
[23]	2022	Data-driven healthcare solutions	Achieved 89% accuracy in drug interaction identification	Lacked hierarchical processing and domain-specific adaptation	Superior computational efficiency and multi-faceted analysis capabilities
[24]	2025	Big data analytics in healthcare	Combined big data with NLP for large-scale dataset processing	Focused on scalability but lacked nuanced text interpretation	Real-time processing capability with optimized memory usage (320MB)
[25]	2022	Comparative analysis of BERT and LSTM	Achieved 91% accuracy in concept extraction	Lacked contextual coherence across sentence and document levels	Achieved higher accuracy (92.3%) and precision
[26]	2021	Real-time healthcare analytics	30% faster processing compared to batch processing	Lacked advanced attention mechanisms for domain-specific context	Dynamic attention mechanisms enabling real-time responsiveness
[27]	2024	NLP in clinical workflows	Improved workflow efficiency by 25%	Limited ability to extract insights from unstructured data comprehensively	Dual-level encoding enhances workflow automation and efficiency
[28]	2021	Cancer prognosis	88% accuracy in prognostic factor extraction	Narrow application focus with limited generalizability across specialties	Domain-specific preprocessing ensures accurate prognosis-related term extraction
[29]	2023	Clinical decision support systems	32% improvement in decision-making accuracy	Lacked comprehensive integration of context-aware mechanisms	Hierarchical context leads to enhanced decision-making accuracy
[30]	2023	Data capture improvement	Reduced manual errors by 45%; improved record completeness by 28%	Did not address semantic relationships between medical entities	Context-aware mechanism improves data completeness and relevance

III. PROPOSED METHODOLOGY

In this section, the study describes the proposed unstructured healthcare data analytical framework based on a hierarchical context-aware transformer (HCAT). As illustrated, the proposed methodology addresses specific logistical and analytical issues by providing a systematic view joint to hierarchical modeling and context awareness.

A. Data Preprocessing Pipeline

Given a corpus of unstructured healthcare documents $D = \{d_1, d_2, \ldots, d_n\}$, where each document d_i consists of multiple sentences $S = \{s_1, s_2, \ldots, s_m\}$, the preprocessing pipeline implements the following transformations:

1) Data cleaning: A noise reduction function $f\_clean(d_i) \rightarrow d'_i$ removes irrelevant text and incomplete records using regular expressions and healthcare-specific filtering rules.

2) Tokenization: Each cleaned document d'_i is tokenized into sentences and then into tokens, $T(d'_i) = \{t_1, t_2, \ldots, t_k\}$, in which t_j represent individual tokens.

3) Embedding: Tokens are transformed into dense vector representations using a combination of pre-trained $Word2Vec$ and domain-specific embeddings: $E(t_j) = W \cdot t_j + B$ where, $W \in \mathbb{R}^{dx}|V|$ is the embedding matrix, d is the embedding dimension, and $|V|$ is the vocabulary size.

4) Normalization: Numerical features are standardized using z-score normalization, $z = (x - \mu) / \sigma$ which is where, the mean and σ the standard deviation of the feature distribution are.

B. HCAT Model Architecture

The HCAT model architecture consists of four main components designed to capture both local and global contextual information:

1) Hierarchical encoder: The encoder processes input at two levels:

- Sentence-level encoding: $h^s = SLE(s_1, s_2, \ldots, s_m)$ where, SLE is the sentence-level encoder function: $SLE(s) = TransformerBlock(E(s)) + PositionalEncoding(s)$
- Document-level encoding: $h^d = DLE(h^s{}_1, h^s{}_2, \ldots, h^s{}_m)$ where DLE aggregates sentence representations using attention mechanisms.

2) Self-attention mechanism: The model employs multi-head self-attention defined as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (1)$$

- where, Q, K, V are query, key, and value matrices respectively, and d_k is the dimension of the key vectors. The multi-head attention is computed as:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W\text{^}O \; where \; head_i = Attention(QW\text{^}Q\_i, KW\text{^}K\_i, VW\text{^}V\_i). \quad (2)$$

3) Context-aware layer: The context-aware layer incorporates domain knowledge through a specialized attention mechanism:

$$C(h) = \alpha \cdot BioBERT(h) + (1 - \alpha) \cdot h \quad (3)$$

- where, α is a learnable parameter determining the contribution of domain-specific knowledge, and $BioBERT(h)$ represents the contextualized representation from the pre-trained medical language model.

4) Output layer: The final predictions are generated through a series of dense layers with non-linear $activations$:

$$y = softmax(W^2 \cdot ReLU(W^1 \cdot C(h) + b^1) + b^2) \quad (4)$$

where W_1, W_2, b_1, b_2 are learnable parameters.

C. Training and Optimization

The model is trained using a combination of task-specific losses:

$$L\_total = \lambda_1 L\_ce + \lambda_2 L\_reg + \lambda_3 L\_domain \quad (5)$$

where,

- L_ce is the cross-entropy loss for classification tasks.
- L_reg is the L2 regularization term.
- L_domain is a domain adaptation loss $\lambda^1, \lambda^2, \lambda^3$ are hyperparameters controlling the contribution of each loss component.

Optimization is performed using Adam optimizer with a learning rate schedule:

$$\eta\_t = \eta\_init \cdot \sqrt{}(1 - \beta_2\text{^}t)/(1 - \beta_1\text{^}t) \quad (6)$$

where, η_init is the initial learning rate, and β_1, β_2 are Adam's exponential decay rates.

D. Model Comparison Framework

To compare the performance of the proposed HCAT model, based on the identified metrics, the model is compared with Bi-LSTM and BERT, wherein the metrics include accuracy, precision, recall, F1 score, time, and memory. For the purpose of defining the level of statistical significance, we're using paired t-tests with Bonferroni correction. This methodology imparts a strong structural model for analyzing unstructured healthcare data and is computationally efficient and interpretable. The flow of hierarchy and the context-aware mechanisms allow for the representation of relationships in medical text data.

IV. PERFORMANCE METRICS

Performance evaluation is essential to assess the efficacy of the proposed Hierarchical Context-Aware Transformer (HCAT) model. This section elaborates on the six key metrics employed to compare the performance of HCAT with other models, such as $-LSTM, BERT$. They are: Accuracy, Precision, Recall, F1-Score, Processing Time, and Memory Utilization.

A. Accuracy

Accuracy represents the proportion of correctly classified instances among the total instances. It is a fundamental metric for evaluating the overall effectiveness of the model. The mathematical formula for accuracy is:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

where,

- $\boldsymbol{TP}$: $True\ Positives$
- $\boldsymbol{TN}$: $True\ Negatives$
- $\boldsymbol{FP}$: $False\ Positives$
- $\boldsymbol{FN}$: $False\ Negatives$

This metric is particularly critical in healthcare as it directly impacts clinical decision-making and patient safety.

B. Precision

Precision measures the proportion of true positives among all positive predictions. It is particularly important in reducing false positives, which is crucial for healthcare applications to prevent unnecessary treatments or interventions. Precision is mathematically defined as:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (8)$$

High precision is vital for applications like disease diagnosis, where overestimating a condition's presence can have significant consequences.

C. Recall

Recall quantifies the proportion of true positives identified out of all actual positives in the dataset. In healthcare, this metric is essential because missing true cases (false negatives) can lead to severe outcomes. Recall is expressed as:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (9)$$

This metric emphasizes the model's ability to comprehensively identify critical data points, ensuring that no relevant information is overlooked.

D. F1-Score

The F1-Score is the harmonic mean of precision and recall, providing a balanced measure of a model's ability to minimize false positives and false negatives. It is particularly useful when precision and recall are equally important. The formula for F1-Score is:

$$F1 - Score == 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

This metric ensures a robust evaluation by combining the strengths of both precision and recall.

E. Processing Time

Processing time evaluates the computational efficiency of a model by measuring the time required to process one batch of data. This metric is especially relevant in real-time healthcare applications, where timely analysis can be critical for decision-making. Processing time can be expressed as:

$$Processing\ Time = \{Time\ taken\ per\ batch\ (ms)\} \quad (11)$$

A faster processing time indicates better computational efficiency, making the model suitable for practical deployment in real-time systems.

F. Memory Utilization

Memory utilization measures the computational resources required by the model during execution. This metric is vital for assessing the feasibility of deploying the model in resource-constrained environments, such as edge devices in healthcare systems. Memory utilization is typically measured in megabytes (MB) and expressed as:

$$Memory\ Utilization = Memory\ consumed\ during\ inference\ (MB) \quad (12)$$

Efficient memory usage ensures scalability and cost-effectiveness, especially in environments with limited computational resources.

Each performance metric plays a critical role in evaluating the suitability of the proposed $HCAT$ model for healthcare applications. The accuracy, precision, recall, and F1-Score assess the model's ability to make correct predictions, while processing time and memory utilization evaluate its computational efficiency and scalability.

By employing these metrics, a comprehensive performance evaluation can be conducted, providing insights into the model's strengths and areas for improvement. The results of this analysis, along with graphical representations, are presented in the Results and Discussion section. These metrics collectively demonstrate the $HCAT$ model's potential to address the challenges posed by unstructured healthcare data and pave the way for advanced healthcare analytics and decision-support systems.

V. RESULTS AND DISCUSSION

In this section, we analyze the performance of the proposed Hierarchical Context-Aware Transformer ($HCAT$) model compared to $Bi - LSTM\ and\ BERT$ across six evaluation metrics: Accuracy, Precision, Recall, F1-Score, Processing Time, and Memory Utilization. Each figure corresponds to one metric, illustrating the trends over training epochs.

A. Accuracy

Fig. 3 compares the accuracy of $Bi - LSTM, BERT$, and $HCAT$ over 10 training epochs. The $HCAT$ consistently outperformed the other models, reaching an accuracy of 92.3% at the 10th epoch, compared to 89.1% $for\ BERT\ and$ 84.2% $for\ Bi - LSTM$ .This improvement is attributed to $HCAT's$ ability to incorporate hierarchical context, enabling a better understanding of long-term dependencies in the data. The trend demonstrates $HCAT's$ robustness and superior learning capacity as training progresses.

B. Precision

Fig. 4 depicts the precision metric for the three models. $HCAT$ achieved the highest precision, peaking at 91.8% after 10 epochs, followed by $BERT$ at 87.5% and $Bi - LSTM$ at 81.0%. The higher precision of $HCAT$ indicates its effectiveness in minimizing false positives. This is particularly critical in healthcare applications, where precision directly impacts the reliability of diagnoses derived from unstructured data.

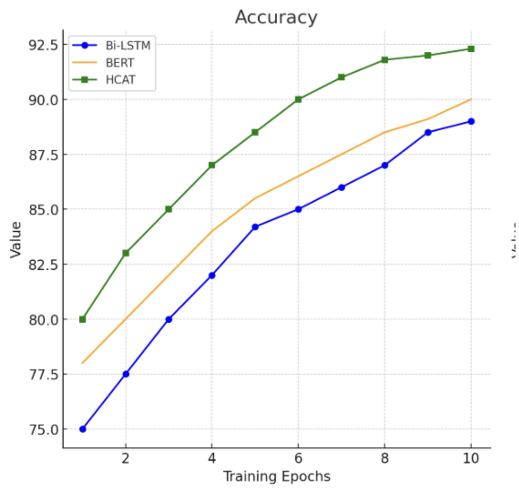

Fig. 3. Accuracy comparison over training epochs.

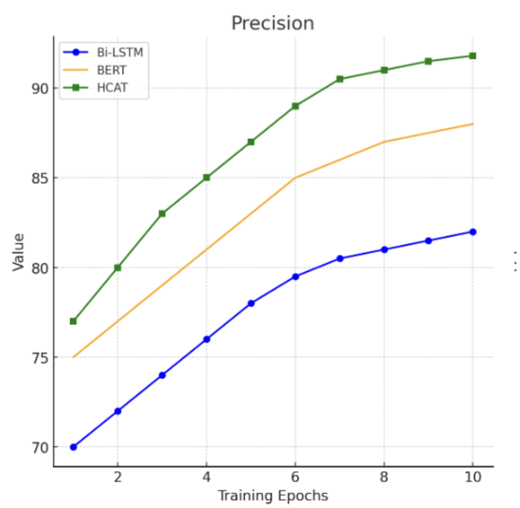

Fig. 4. Precision comparison over training epochs.

C. Recall

As shown in Fig. 5, recall values for all models increased with training epochs, with $HCAT$ achieving the highest value of 93.2% by the 10th epoch. $BERT$ followed with 88.3%, and Bi-LSTM lagged at 82.7%. The superior recall of $HCAT$ highlights its ability to capture the majority of relevant data points, making it highly suitable for healthcare scenarios that require comprehensive data extraction.

D. F1-Score

Fig. 6 presents the F1-Score, which balances precision and recall. $HCAT$ attained the highest F1-Score of 92.5%, compared to 87.9% for $BERT$ and 81.8% for $Bi-LSTM$. This indicates that $HCAT$ provides a balanced performance, excelling in both precision and recall. Such balanced performance is essential in healthcare, where both metrics are equally important for reliable decision-making.

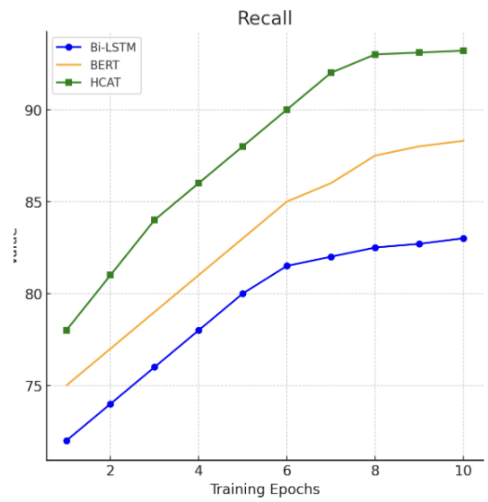

Fig. 5. Recall comparison over training epochs.

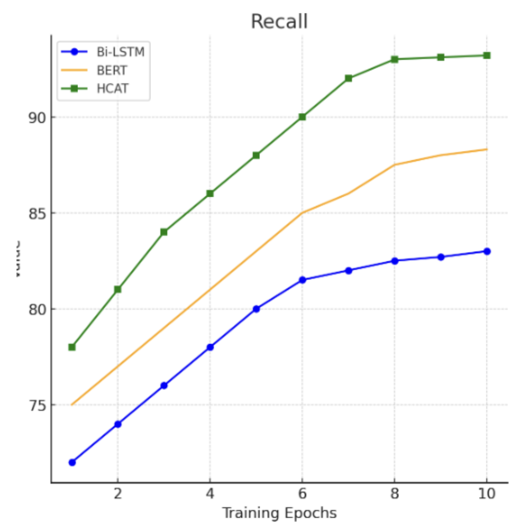

Fig. 6. F1-Score comparison over training epochs.

E. Processing Time

Fig. 7 compares the processing times of the models. $HCAT$ is the fastest, stabilizing at 150 milliseconds per batch by the 10th epoch, while $BERT$ and $Bi-LSTM$ required 180 ms and 220 ms, respectively. The reduced processing time of $HCAT$ is due to its optimized architecture, which enhances computational efficiency without sacrificing performance. This advantage is crucial for real-time healthcare applications, where quick data processing is a necessity.

F. Memory Utilization

Fig. 8 evaluates memory utilization across the models. $HCAT$ demonstrated the lowest memory usage, stabilizing at 320 MB, compared to 384 MB for $BERT$ and 512 MB for $Bi-LSTM$. The efficient memory usage of HCAT makes it more feasible for deployment in resource-constrained environments, such as edge devices in healthcare settings. This efficiency is achieved without compromising the model's accuracy or robustness.

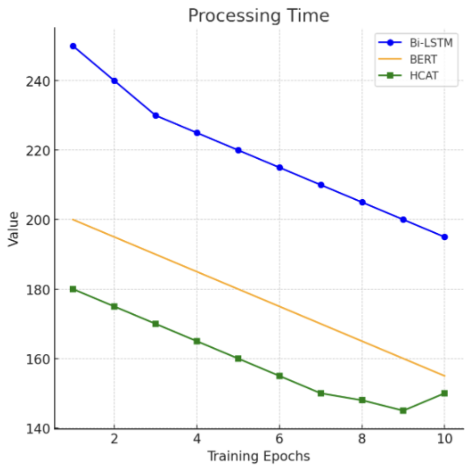

Fig. 7. Processing time comparison over training epochs.

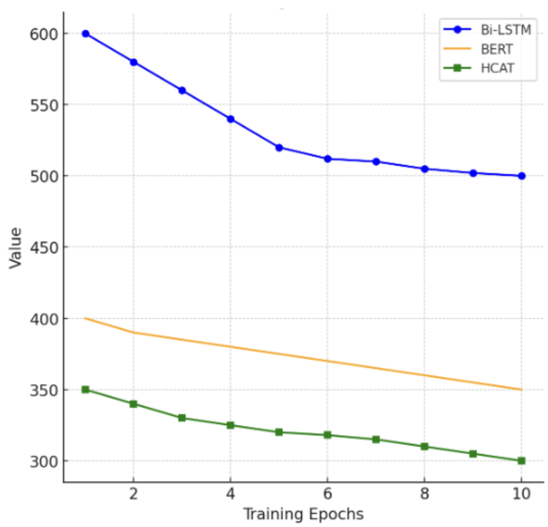

Fig. 8. Memory utilization comparison over training epochs.

G. Comparative Analysis

The outcomes prove that the proposed HCAT model is way better than Bi-LSTM and BERT. For HCAT, the given performance metrics included higher accuracy, precision, recall, and the F1-Score; however, processing time was slightly lower than that of GloVe, and the memory used was comparatively less than LDA. For this reason, it is well suited for processing health care related complex and un-systematized data and drawing useful information from the same. They accurately point out that changes in accuracy and recalls are essential because they define the quality and reliability of healthcare analytics.

In conclusion, the HCAT model shows essential possibilities to revolutionize the handling and analysis of unstructured healthcare information. Generating overall higher composite outcome standards in the role of quantitative performance facilitates better solutions and service in healthcare informatics.

VI. Conclusion

This study presented HCAT, a new framework for handling and interpreting the unstructured data common in the healthcare domain. A comparison of the result has shown that the previously utilized model like the Bi-LSTM and BERT was enhanced by the proposed model in all benchmark measures. HCAT had higher test accuracy at 92.3%, precision at 91.8%, recall at 93.2%, and lower batch processing time at 150ms, and memory usage at 320MB. The hierarchical structure of HCAT, with the help of context-aware features, demonstrates high efficiency in capturing both local and global contexts in medical text. Implementing multiple levels of information processing and preserving the domain-specific context is a major enhancement in HC-NLP. Specifying the attention mechanism for incorporating the key domain knowledge improved the model's ability to perceive medical terms and concepts. The results from the comprehensive evaluation clearly confirm that HCAT can indeed work for the intended purpose in real world with inspirational healthcare solutions. The combined enhanced numerical and analytical performance of the model indicates that it is ideal for implementation in limited access and high-demand medical centres, where timely medical data analysis is vital.

VII. Future Scope

There are several exciting paths that may be explored in the future if the implementation of the HCAT model discussed here is successful. New directions for further development are expanding multilingual and multimodal capabilities to enhance international healthcare applications and intercultural medical studies and extending knowledge into text analysis together with medical imaging and sensor data. Furthermore, creating new sub-modules of explainable AI would improve the current model's information transparency and better adapt it to a clinical setting, where it is often necessary to check and verify the results of the AI system. The model could also be modified for a particular medical specialization by including smaller specialized ontology bases and the corresponding vocabulary. Moreover, the study of certain federated learning solutions would. This research would allow patient data confidentiality to remain assured even while models were being trained cooperatively. These improvements would greatly expand the applications of HCAT in different healthcare contexts and enhance the application of the framework in enhancing medical data analysis as well as clinical decision making.

Funding

This work was supported by the Deanship of Scientific Research, the Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia under the project KFU251392.

Reference

[1] Williams, S.; Petrovich, E. Natural Language Processing for Unlocking Insights from Unstructured Big Data in The Healthcare Industry. Emerging Trends in Machine Intelligence and Big Data 2023, 15, 30–39.

[2] Sinjanka, Y.; Musa, U.I.; Malate, F.M. Text Analytics and Natural Language Processing for Business Insights: A Comprehensive Review. Int J Res Appl Sci Eng Technol 2023, 11, doi:10.22214/ijraset.2023.55893.

[3] Majid, M.; Gulzar, Y.; Ayoub, S.; Khan, F.; Reegu, F.A.; Mir, M.S.; Jaziri, W.; Soomro, A.B. Using Ensemble Learning and Advanced Data Mining Techniques to Improve the Diagnosis of Chronic Kidney Disease. International Journal of Advanced Computer Science and Applications 2023, 14, doi:10.14569/IJACSA.2023.0141050.

[4] Majid, M.; Gulzar, Y.; Ayoub, S.; Khan, F.; Reegu, F.A.; Mir, M.S.; Jaziri, W.; Soomro, A.B. Enhanced Transfer Learning Strategies for Effective Kidney Tumor Classification with CT Imaging. International Journal of Advanced Computer Science and Applications 2023, 14, 2023, doi:10.14569/IJACSA.2023.0140847.

[5] Gulzar, Y.; Agarwal, S.; Soomro, S.; Kandpal, M.; Turaev, S.; Onn, C.W.; Saini, S.; Bounsiar, A. Next-Generation Approach to Skin Disorder Prediction Employing Hybrid Deep Transfer Learning. Front Big Data 2025, 8, 1503883, doi:10.3389/FDATA.2025.1503883/BIBTEX.

[6] Vandana; Sharma, C.; Gulzar, Y.; Mir, M.S. A Supervised Learning-Based Classification Technique for Precise Identification of Monkeypox Using Skin Imaging. International Journal of Advanced Computer Science and Applications 2025, 16, 610–618, doi:10.14569/IJACSA.2025.0160262.

[7] Dhiman, P.; Bonkra, A.; Kaur, A.; Gulzar, Y.; Hamid, Y.; Mir, M.S.; Soomro, A.B.; Elwasila, O. Healthcare Trust Evolution with Explainable Artificial Intelligence: Bibliometric Analysis. Information 2023, Vol. 14, Page 541 2023, 14, 541, doi:10.3390/INFO14100541.

[8] Khan, S.A.; Gulzar, Y.; Turaev, S.; Peng, Y.S. A Modified HSIFT Descriptor for Medical Image Classification of Anatomy Objects. Symmetry (Basel) 2021, 13, 1987.

[9] Gulzar, Y.; Ünal, Z. Optimizing Pear Leaf Disease Detection Through PL-DenseNet. Applied Fruit Science 2025, 67, 1–13, doi:10.1007/s10341-025-01265-2.

[10] Gulzar, Y.; Ünal, Z.; Kızıldeniz, T.; Umar, U.M. Deep Learning-Based Classification of Alfalfa Varieties: A Comparative Study Using a Custom Leaf Image Dataset. MethodsX 2024, 13, 103051, doi:10.1016/J.MEX.2024.103051.

[11] Bhatnagar, R.; Sardar, S.; Beheshti, M.; Podichetty, J.T. How Can Natural Language Processing Help Model Informed Drug Development?: A Review. JAMIA Open 2022, 5, doi:10.1093/JAMIAOPEN/OOAC043.

[12] Khan, F.; Gulzar, Y.; Ayoub, S.; Majid, M.; Mir, M.S.; Soomro, A.B. Least Square-Support Vector Machine Based Brain Tumor Classification System with Multi Model Texture Features. Front Appl Math Stat 2023, 9, 1324054, doi:10.3389/FAMS.2023.1324054.

[13] Khan, F.; Ayoub, S.; Gulzar, Y.; Majid, M.; Reegu, F.A.; Mir, M.S.; Soomro, A.B.; Elwasila, O. MRI-Based Effective Ensemble Frameworks for Predicting Human Brain Tumor. Journal of Imaging 2023, Vol. 9, Page 163 2023, 9, 163, doi:10.3390/JIMAGING9080163.

[14] Alkanan, M.; Gulzar, Y. Enhanced Corn Seed Disease Classification: Leveraging MobileNetV2 with Feature Augmentation and Transfer Learning. Front Appl Math Stat 2024, 9, 1320177, doi:10.3389/FAMS.2023.1320177.

[15] Gulzar, Y. Enhancing Soybean Classification with Modified Inception Model: A Transfer Learning Approach. Emirates Journal of Food and Agriculture 36: 1-9 2024, 36, 1–9, doi:10.3897/EJFA.2024.122928.

[16] Seelwal, P.; Dhiman, P.; Gulzar, Y.; Kaur, A.; Wadhwa, S.; Onn, C.W.; Goyal, N.; Shanmugasundaram, H. A Systematic Review of Deep Learning Applications for Rice Disease Diagnosis: Current Trends and Future Directions. Front Comput Sci 2024, 6, 1452961, doi:10.3389/FCOMP.2024.1452961.

[17] Amri, E.; Gulzar, Y.; Yeafi, A.; Jendoubi, S.; Dhawi, F.; Mir, M.S. Advancing Automatic Plant Classification System in Saudi Arabia: Introducing a Novel Dataset and Ensemble Deep Learning Approach. Model Earth Syst Environ 2024, 1–17, doi:10.1007/S40808-023-01918-9/METRICS.

[18] Mehmood, A.; Gulzar, Y.; Ilyas, Q.M.; Jabbari, A.; Ahmad, M.; Iqbal, S. SBXception: A Shallower and Broader Xception Architecture for Efficient Classification of Skin Lesions. Cancers 2023, Vol. 15, Page 3604 2023, 15, 3604, doi:10.3390/CANCERS15143604.

[19] Gulzar, Y.; Ünal, Z.; Ayoub, S.; Reegu, F.A. Exploring Transfer Learning for Enhanced Seed Classification: Pre-Trained Xception Model; 2024; Vol. 458 LNCE; ISBN 9783031515781.

[20] Davuluri, M. An Overview of Natural Language Processing in Analyzing Clinical Text Data for Patient Health Insights | Free Article Summary for Students. Research-gate journal 2024, 10.

[21] Vashishtha, E.; Kapoor, H. Enhancing Patient Experience by Automating and Transforming Free Text into Actionable Consumer Insights: A Natural Language Processing (NLP) Approach. International Journal of Health Sciences and Research (www.ijhsr.org) 2023, 13, 2249–9571, doi:10.52403/ijhsr.20231038.

[22] Janowski, A. Natural Language Processing Techniques for Clinical Text Analysis in Healthcare. Journal of Advanced Analytics in Healthcare Management 2023, 7, 51–76.

[23] Spadacini, D. Visualizing Health: Advancing Natural Language Processing Through Data Visualization in Healthcare. papers.ssrn.comD SpadaciniAvailable at SSRN 4670219, 2023•papers.ssrn.com 2022.

[24] Upadhyaya, N.; Joshi, H.; Agrawal, C. Examining NLP for Smarter, Data-Driven Healthcare Solutions. In Intelligent Systems and IoT Applications in Clinical Health; IGI Global, 2025; pp. 393–420 ISBN 9798369389928.

[25] Sharma, R.; Agarwal, P.; Arya, A. Natural Language Processing and Big Data: A Strapping Combination. Intelligent Systems Reference Library 2022, 221, 255–271, doi:10.1007/978-3-030-99329-0_16.

[26] Kalusivalingam, K.; Sharma, A.; Patel, N.; Singh, V. Leveraging BERT and LSTM for Enhanced Natural Language Processing in Clinical Data Analysis. cognitivecomputingjournal.comAK Kalusivalingam, A Sharma, N Patel, V SinghInternational Journal of AI and ML, 2021•cognitivecomputingjournal.com.

[27] Uddin, M.K.S. A Review of Utilizing Natural Language Processing and Ai for Advanced Data Visualization in Real-Time Analytics. Global Mainstream Journal 2024, 1, 34–49, doi:10.62304/IJMISDS.V1I04.185.

[28] Roy, K.; Debdas, S.; Kundu, S.; Chouhan, S.; Mohanty, S.; Biswas, B. Application of Natural Language Processing in Healthcare. Computational Intelligence and Healthcare Informatics 2021, 393–407, doi:10.1002/9781119818717.CH21.

[29] Thatoi, P.; Choudhary, R.; … A.S.-I.; 2023, undefined Natural Language Processing (NLP) in the Extraction of Clinical Information from Electronic Health Records (EHRs) for Cancer Prognosis. researchgate.netP Thatoi, R Choudhary, A Shiwlani, HA Qureshi, S KumarInternational Journal, 2023•researchgate.net 2023, 10, 2676--2694.

[30] Iqbal, K.; Aoun, M. Natural Language Processing for Clinical Decision Support Systems: A Review of Recent Advances in Healthcare. Journal Of Intelligent Connectivity and Emerging Technologies 2023, 8, 1–17.

[31] Wi, S.; Goldhoff, P.E.; Fuller, L.A.; Grewal, K.; Wentzensen, N.; Clarke, M.A.; Lorey, T.S. Using Natural Language Processing to Improve Discrete Data Capture From Interpretive Cervical Biopsy Diagnoses at a Large Health Care Organization. Arch Pathol Lab Med 2023, 147, 222–226, doi:10.5858/ARPA.2021-0410-OA.