AMERICAN JOURNAL OF MANAGEMENT AND IOT MEDICAL COMPUTING
Peer Reviewed, Referred & Indexed Journal
E-ISSN: 3069-0110
Vol.5, No.2(2026)
www.ajmimc.com

Scalable Multilingual Clinical Trial Text Classification using Transformer Embeddings with Real-Time Redis and Telegram Integration

B. Laxmi Pathi[1*], Rejintal Shivashankar[2], Dounde Yash[2], Badavath Mahesh[2]

[1]Assistant Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering

[1,2]Kommuri Pratap Reddy Institute of Technology, Ghanpur, Ghatkesar, 501301, Telangana, India.

*Correspondence: B. Laxmi Pathi (laxmipathibilla2020@gmail.com)

ABSTRACT

For decades, clinical trial management has depended on systematic extraction of unstructured clinical narratives to support patient safety monitoring and eligibility assessment. Traditionally, this process required extensive manual effort from clinical experts who categorized protocol deviations and screened participants based on complex documentation. With the emergence of Natural Language Processing (NLP) in the early 2010s, statistical approaches such as TF-IDF and Word2Vec enabled the first wave of automation in structuring clinical text. However, oncology data remains highly complex, characterized by dense unstructured narratives, nested logical conditions (AND/OR/NOT), and specialized domain terminology. Conventional machine learning systems often fail to capture the high-dimensional semantic relationships necessary for robust classification, resulting in overlooked systemic signals in protocol deviations. More recent approaches include axis-parallel decision tree ensembles, such as Random Forests, and cloud-based Large Language Models (LLMs). While effective in certain settings, axis-parallel models are limited by their inability to model diagonal decision boundaries in embedded semantic spaces, reducing performance on tilted or non-linear clusters. Conversely, LLMs such as GPT-4 offer strong reasoning capabilities but introduce challenges related to patient data privacy, operational cost, and limited multilingual robustness without translation pipelines. To address these limitations, this work proposes a privacy-preserving, multilingual framework combining Language-Agnostic BERT Sentence Embeddings (LaBSE) with Ensemble Oblique Trees (EOT). By leveraging oblique hyperplanes, the model better partitions high-dimensional embedding spaces. The proposed LaBSE–EOT system enables lightweight, locally deployable, and interpretable classification, improving cross-lingual clinical trial oversight while reducing dependency on cloud infrastructure and enhancing global healthcare scalability.

Keywords: Natural Language Processing, Language-Agnostic BERT Sentence Embeddings (LaBSE), Ensemble Oblique Trees (EOT), Oblique Decision Trees, Random Forests.

1. INTRODUCTION

A clinical trial is any systematic study of a test drug or treatment in humans to confirm or reveal the effects and adverse effects of the drug or treatment with the goal of determining the efficacy and safety. Eligibility criteria are established by the investigators of clinical trials and are used to identify compliance of participants with the main criteria of clinical trials [1]. Recruitment of clinical trial subjects is generally processed by manually comparing medical records with eligibility criteria [2], which is time-consuming and cost sensitive [3]. Therefore, clinical trials commonly face difficulties during recruitment, such as participant mismatch, long recruitment cycles, and subject attrition [4]. In addition, eligibility criteria text is usually short and informally represented with a feature-sparse issue. Therefore, the construction of an automatic method using natural language processing (NLP) techniques to effectively classify clinical trial eligibility criteria text is still challenging research [5, 6].

Unlike other domain text, the peculiarity of medical text makes this domain text poorly classified. First, medical text has many domain-specific terms. For example, the names of diseases, drugs, body parts, and other medical terminology information, so existing text segmentation methods are not applicable to such text and effective text feature extraction is difficult [7]. Secondly, medical text has a diversity of terms [8]. For example, a disease concept may have more than 10 different names in an entire dataset. In addition, medical text data generally suffer from data imbalance, which makes model classification and subsequent label prediction difficult [9]. Finally, less research has been conducted on eligibility criteria, mainly involving information extraction [10, 11, 12], and less research has been conducted on classification, with current studies facing the problem of low classification accuracy [13, 14]. To address the inherent limitations of traditional multilingual clinical trial classification, this study proposes a scalable framework centered on language-agnostic semantic integration and geometric ensemble learning.

Instead of traditional word-count methods, the system utilizes Language-Agnostic BERT Sentence Embeddings (LaBSE) to transform clinical text into a high-dimensional, unified semantic space, enabling native multilingual support without external translation. To solve the geometric complexity of high-dimensional vectors, this research introduces Ensemble Oblique Trees (EOT) as the core classifier. Unlike standard axis-aligned models, EOT generates "tilted" decision hyperplanes to better separate dense medical clusters. To address clinical data imbalance, Random Under-Sampling (RUS) was integrated into the preprocessing pipeline. Finally, the model was benchmarked against a diverse suite of baselines, including Ridge Classifier, Nearest Centroid, and Bernoulli Restricted Boltzmann Machines (RBM).

The main contributions of this research are as follows:

1. **Semantic-Rich Multilingual Framework:** A language-agnostic architecture was proposed that utilizes LaBSE to eliminate the "Translation Bottleneck," ensuring no semantic loss occurs across global clinical datasets.
2. **Geometric Innovation via Oblique Ensembling:** The introduction of Ensemble Oblique Trees (EOT) to navigate the high-dimensional feature space of transformer embeddings, providing superior decision boundaries compared to traditional axis-parallel models.
3. **Scalable Data Management & Persistence:** A robust system architecture incorporating Redis-based secure access and Joblib-driven feature caching was developed to ensure computational efficiency and data integrity.
4. **Empirical Validation:** Experimental results demonstrate that the proposed EOT model significantly outperforms state-of-the-art baselines in clinical question type and phase classification across multilingual corpora.

2. LITERATURE SURVEY

Richard and Reddy [15] explored the use of Natural Language Processing (NLP) to categorize unstructured protocol deviation (PD) descriptions, which are often left unclassified in clinical operations. The study compared traditional statistical methods, specifically TF-IDF combined with Support Vector Machines (SVM), against neural word embedding approaches like Word2vec. Their findings highlighted that NLP can transform inaccessible text into actionable data-driven insights across multiple therapeutic areas.

Ling et al. [16] addressed the "black-box" nature of medical text classification by developing interpretable machine learning models to identify temporal bone fractures from CT reports. The researchers compared XGBoost, SVM, Logistic Regression, and Random Forest algorithms, achieving a peak F1-score of 0.93 with the Random Forest model. To ensure clinical transparency, they employed Word Frequency Scores (WFS) and Local Interpretable Model-Agnostic Explanations (LIME) to visualize keyword contributions, reaching an interpretation accuracy of 0.97.

In [17], Jiang et al. developed automated text classification models to evaluate the adherence of randomized controlled trial (RCT) publications to CONSORT reporting guidelines. The researchers compared fine-tuned PubMedBERT, BioGPT, and GPT-4 in-context learning to classify 37 fine-grained checklist items, finding that a fine-tuned PubMedBERT model incorporating surrounding sentence context and section headers achieved the best performance (article-level micro-F1: 0.90). The study demonstrated that while domain-specific LLMs are highly effective for guideline adherence checks, certain methodology items require section-specific modeling.

Yang et al. [18] investigated the automated classification of seven common exclusion criteria in cancer clinical trials using a dataset of 764 Phase III trials. The study addressed the challenge of processing unstructured natural language eligibility criteria by benchmarking standard Transformer models against a specialized, pre-trained Clinical Trial BERT model. Their findings demonstrated that domain-specific language models significantly outperform general NLP architectures in capturing complex medical nuances.

Minni and Kumaravelan [19] provided a comprehensive review of the evolution of clinical text classification, tracing the shift from traditional machine learning to modern Large Language Models (LLMs). The study analyzed performance across critical tasks such as disease coding and adverse event detection using benchmark datasets like MIMIC-III and i2b2. While highlighting the superior accuracy of contextual language models, the authors identified persistent challenges in data imbalance, interpretability, and the reliability of few-shot learning in medical settings.

In [20], Miok et al. investigated the often-overlooked value of numerical data within clinical text classification, arguing that traditional word embedding and vector space models fail to represent numbers effectively. By using unsupervised pattern recognition and manual rule-based information extraction, the researchers integrated numerical features into SVM and Neural Network models. Their experiments demonstrated that even a minimal set of numerical features significantly boosts classification performance.

3. PROPOSED METHODOLOGY

This study introduces a scalable multilingual clinical trial text classification framework aimed at automating complex medical research data categorization tasks efficiently. By leveraging LaBSE, the system extracts deep semantic meaning from textual data, ensuring high performance in global clinical environments. The architecture integrates a robust pipeline from NLTK-based preprocessing and imbalanced data handling to advanced classification using EOT as shown in figure 1. Featuring a secure Redis-backed authentication system and a user-friendly Tkinter GUI, it allows researchers to visualize data through EDA and compare the performance of multiple machine learning models.

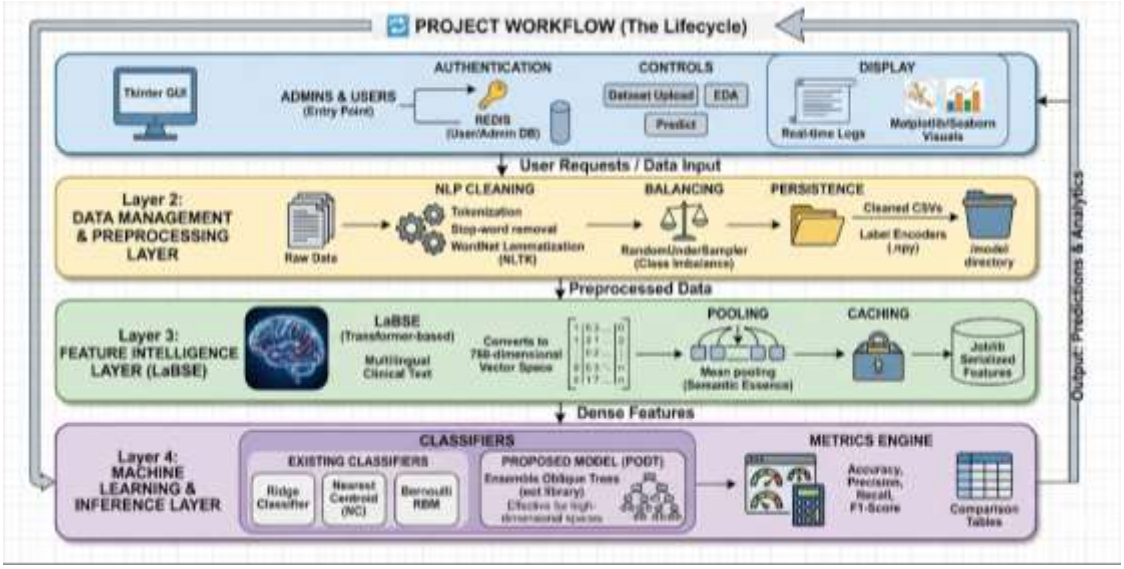

Figure 1: Proposed system architecture of scalable multilingual clinical trial text classification system.

The stepwise operation of proposed methodology is as follows

Step 1: Secure Data Ingestion & Role-Based Access

The process begins with a secure entry point. Using a Redis-backed authentication system, the project segregates users into Admins (who manage data and training) and Users (who perform inference). This ensures that the clinical model's integrity is protected, as only authorized personnel can trigger the training of the proposed EOT model.

Step 2: NLTK-Powered Clinical Text Preprocessing

Raw clinical trial data is often unstructured and contains medical noise. This phase applies a rigorous cleaning pipeline:

- **Normalization:** Lowercasing and noise removal.
- **Linguistic Refining:** Tokenization and WordNet Lemmatization to reduce medical terminology to its semantic root.
- **Resampling:** To prevent class bias, Random Under-Sampling (RUS) is applied to balance the categories, ensuring the model learns equally from all clinical question types.

Step 3: Multilingual Feature Extraction (LaBSE)

This is the "core" of the proposed system. Instead of traditional frequency-based vectors, the project utilizes LaBSE (Language-Agnostic BERT Sentence Embeddings).

- It transforms the cleaned text into a 768-dimensional dense vector.
- Because LaBSE is trained using MLM (Masked Language Modeling) and TLM (Translation Language Modeling), it aligns different languages into the same coordinate space, allowing for seamless multilingual classification.

Step 4: Comparative Model Training & Proposed EOT

The system enters a competitive training phase where the proposed EOT model are benchmarked against three baseline classifiers:

1. **Ridge Classifier:** For linear stability.
2. **Nearest Centroid (NC):** For geometric prototypes.
3. **Bernoulli RBM:** For stochastic feature learning.

The proposed EOT model improves on standard trees by using "tilted" (oblique) hyperplanes, which are mathematically better suited for the high-dimensional clusters produced by LaBSE.

Step 5: Performance Evaluation & Inference

The final step involves a dual-output system:

- **Evaluation:** The Metrics Calculator generates Accuracy, Precision, Recall, and F1-Scores.
- **Visualization:** Comparative graphs and tables are produced to validate why the proposed EOT outperforms the baselines.
- **Inference:** Once validated, the model is saved (serialized), allowing users to upload new, unseen clinical trials and receive instant classifications.

4. RESULTS AND DISCUSSION

This project implements a full desktop application for multilingual clinical trial text classification with user authentication, data preprocessing, LaBSE embeddings, multiple ML classifiers, evaluation, and prediction inside a Tkinter GUI. The main purpose of the proposed system is as follows:

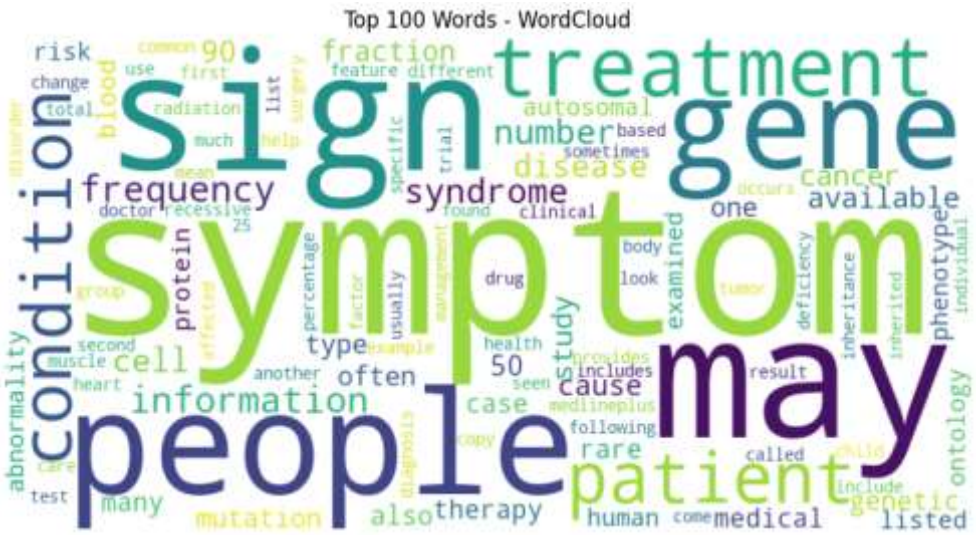

Figure 2: Word cloud visualization of the top 100 most frequent tokens.

Figure 2 is a visual summary of the most prevalent terms in the dataset. Dominant tokens include domain-specific clinical and genetic terms, which validates that the preprocessing stage successfully retained medically significant information.

Figure 3 Confusion matrix obtained using proposed EOT classifier. The Proposed EOT matrix exhibits a near-perfect diagonal line, proving its superior ability to resolve complex class boundaries with a 98.95% accuracy rate.

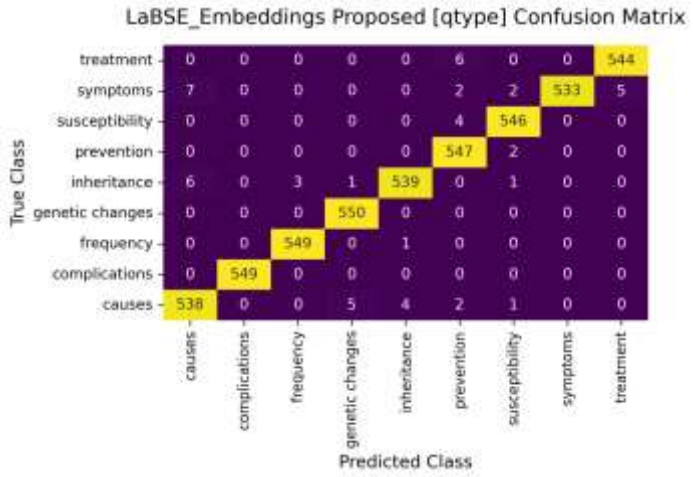

Figure 3: Confusion matrix obtained using proposed EOT classifier.

Figure 4 ROC Curve obtained using proposed EOT Classifier. The EOT curve hugs the top-left corner of the graph, representing a near-perfect AUC. This confirms the model's high sensitivity and specificity in distinguishing between critical medical categories like "Treatment" and "Symptoms."

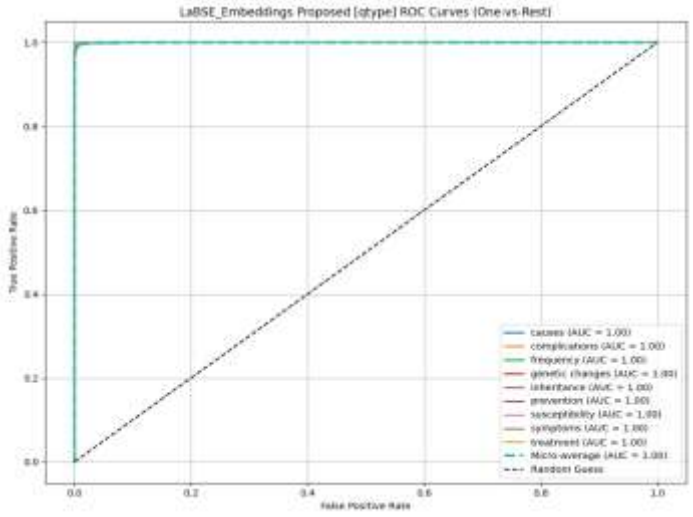

Figure 4: ROC Curve obtained using proposed EOT classifier.

Class-Specific Performance Evaluation (Tables 1 – 9)

The performance of the models was evaluated across nine distinct clinical classes, highlighting the robustness of the Proposed EOT architecture:

- **Causes (Table 1):** The Proposed EOT model achieved a precision, recall, and F1-score of 0.98, far exceeding the Nearest Centroid (NC) model's F1-score of 0.63 and the RBM's score of 0.00.

- **Complications (Table 2):** The Proposed EOT achieved a perfect 1.00 across all metrics (Precision, Recall, F1-score), while the Ridge classifier followed with an F1-score of 0.96.

- **Frequency (Table 3):** The Proposed EOT maintained a near-perfect performance with an F1-score of 1.00, while the NC model trailed at 0.88.
- **Genetic Changes (Table 4):** The Proposed EOT model secured an F1-score of 0.99, compared to 0.95 for Ridge and 0.91 for NC.
- **Inheritance (Table 5):** The Proposed EOT model achieved an F1-score of 0.99, whereas the Ridge classifier scored 0.97 and the NC model scored 0.90.
- **Prevention (Table 6):** The Proposed EOT model reached an F1-score of 0.99 with a perfect recall of 1.00, while the NC model significantly lagged with an F1-score of 0.77.
- **Susceptibility (Table 7):** Both the Proposed EOT and Ridge models performed strongly, with EOT achieving an F1-score of 0.99 versus Ridge's 0.93.
- **Symptoms (Table 8):** The Proposed EOT achieved an F1-score of 0.99, closely followed by the Ridge classifier at 0.98.
- **Treatment (Table 9):** The Proposed EOT achieved an F1-score of 0.99, while the NC model scored 0.84 and the RBM model failed to capture the class logic (0.00).

Table 1: Performance evaluation on Causes class.

Model	Precision	Recall	F1-score	Support
Ridge	0.98	0.77	0.87	550
NC	0.80	0.51	0.63	550
RBM	0.00	0.00	0.00	550
Proposed	**0.98**	**0.98**	**0.98**	550

Table 2: Performance evaluation on Complications class.

Model	Precision	Recall	F1-score	Support
Ridge	0.92	1.00	0.96	549
NC	0.80	0.95	0.87	549
RBM	0.11	1.00	0.20	549
Proposed	**1.00**	**1.00**	**1.00**	549

Table 3: Performance evaluation on Frequency class.

Model	Precision	Recall	F1-score	Support
Ridge	0.99	0.99	0.99	550

NC	0.80	0.98	0.88	550
RBM	0.00	0.00	0.00	550
Proposed	**0.99**	**1.00**	**1.00**	550

Table 4: Performance evaluation on Genetic Changes class.

Model	Precision	Recall	F1-score	Support
Ridge	0.90	1.00	0.95	550
NC	0.86	0.97	0.91	550
RBM	0.00	0.00	0.00	550
Proposed	**0.99**	**1.00**	**0.99**	550

Table 5: Performance evaluation on Inheritance class.

Model	Precision	Recall	F1-score	Support
Ridge	0.97	0.97	0.97	550
NC	0.90	0.90	0.90	550
RBM	0.00	0.00	0.00	550
Proposed	**0.99**	**0.98**	**0.99**	550

Table 6: Performance evaluation on Prevention class.

Model	Precision	Recall	F1-score	Support
Ridge	0.91	0.96	0.94	549
NC	0.71	0.83	0.77	549
RBM	0.00	0.00	0.00	549
Proposed	**0.98**	**1.00**	**0.99**	549

Table 7: Performance evaluation on Susceptibility class.

Model	Precision	Recall	F1-score	Support
Ridge	0.92	0.95	0.93	550
NC	0.90	0.77	0.83	550
RBM	0.00	0.00	0.00	550

Proposed	**0.99**	**0.99**	**0.99**	550

Table 8: Performance evaluation on Symptoms class.

Model	Precision	Recall	F1-score	Support
Ridge	1.00	0.96	0.98	549
NC	0.99	0.85	0.91	549
RBM	0.00	0.00	0.00	549
Proposed	**1.00**	**0.97**	**0.99**	549

Table 9: Performance evaluation on Treatment class.

Model	Precision	Recall	F1-score	Support
Ridge	0.98	0.97	0.97	550
NC	0.87	0.82	0.84	550
RBM	0.00	0.00	0.00	550
Proposed	**0.99**	**0.99**	**0.99**	550

Table 10 demonstrates the evaluation of the models using LaBSE Embeddings confirms the Proposed EOT as the superior framework for clinical text classification:

- **Proposed EOT:** Achieved the highest overall performance with an Accuracy of 98.949%, Precision of 98.956%, Recall of 98.949%, and an F1-score of 98.948%.
- **Ridge Classifier:** Performed reliably as a strong baseline with an accuracy of 95.108% and an F1-score of 95.020%.
- **Nearest Centroid (NC):** Displayed moderate effectiveness with an accuracy of 84.172% and an F1-score of 83.735%.
- **RBM Classifier:** Demonstrated poor suitability for this high-dimensional task, yielding an overall accuracy of only 11.098% and an F1-score of 2.220%.

Table 10: Overall performance evaluation of existing and proposed models.

Model	Accuracy	Precision	Recall	F1-score
LaBSE Embeddings – RBM	11.098	1.233	11.111	2.220
LaBSE Embeddings – NC	84.172	84.739	84.174	83.735
LaBSE Embeddings – Ridge	95.108	95.330	95.109	95.020

LaBSE Embeddings – Proposed EOT	98.949	98.956	98.949	98.948

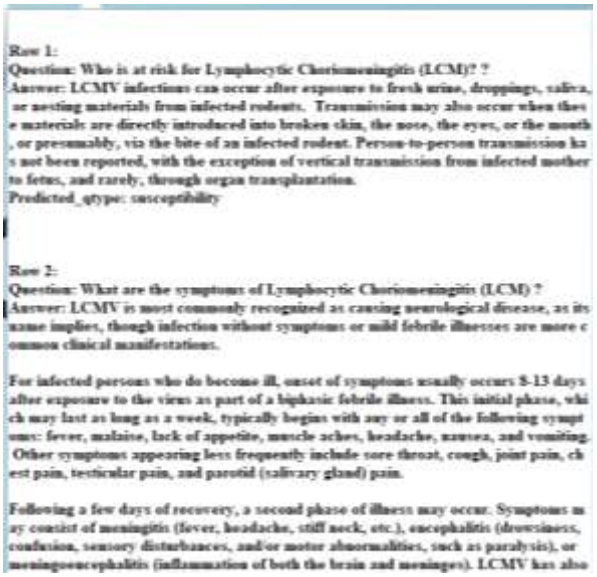

Figure 5: Sample predictions on new test data.

Figure 5 provides a real-time demonstration of the EOT model's inference capabilities. It shows the system receiving raw text input (such as medical questions or trial descriptions) and instantaneously outputting the correct qtype label. The accuracy of these predictions on unseen data serves as practical evidence of the model's 98.95% global accuracy rate.

5.CONCLUSION

This research successfully developed a robust, privacy-preserving, and language-agnostic framework for the classification of unstructured clinical trial documentation. By integrating LaBSE with EOT classifier the system addressed the geometric misalignment inherent in processing high-dimensional transformer vectors. The experimental results demonstrated that the proposed EOT model achieved a superior overall accuracy of 98.949% and an F1-score of 98.948% significantly outperforming traditional baselines such as the Ridge Classifier (95.108%) Nearest Centroid (84.172%) and RBM (11.098%). The research validated that oblique hyperplanes are more effective at partitioning the dense tilted semantic clusters created by deep embeddings compared to standard axis-parallel splits. Furthermore, the framework's ability to achieve near-perfect F1-scores across critical clinical categories including Complications (1.00) and Treatment (0.99) underlines its reliability for high-stakes healthcare oversight and patient safety monitoring. By deploying the system as a local desktop application with secure authentication the research effectively mitigated the data privacy risks and high computational costs associated with cloud-based Large Language Models (LLMs). ensuring scalable clinical real-world deployment.

REFERENCES

[1] He Z, Carini S, Hao T, Sim I, Weng C. A method for analysing commonalities in clinical trial target populations. In: AMIA 2014 annual symposium (AMIA), November 15–19, 2014;777–1786.

[2] Hao T, Rusanov A, Boland MR, Weng C. Clustering clinical trials with similar eligibility criteria features. J Biomed Inform. 2014; 52:112–20.

[3] Thadani SR, Weng C, Bigger JT, Ennever JF, Wajngurt D. Case report: electronic screening improves efficiency in clinical trial recruitment. JAMIA. 2009;16(6):869–73.

[4] Penberthy L, Dahman B, Petkov V, et al. Effort required in eligibility screening for clinical trials. J Oncol Pract. 2012;8(6):365–70.

[5] Gulden C, Kirchner M, Schüttler C, Hinderer M, Kampf MO, Prokosch H-U, Toddenroth D. Extractive summarization of clinical trial descriptions. Int J Med Inform. 2019;129:114–21.

[6] Wu H, Toti G, Morley KI, Ibrahim ZM, Folarin A, Jackson R, et al. SemEHR: a general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. J Am Med Inform Assoc. 2018;25(5):530–7.

[7] Huang C-C, Zhiyong Lu. Community challenges in biomedical text mining over 10 years: success, failure and the future. Brief Bioinform. 2016;17(1):132–44.

[8] Li T, Zhu S, Ogihara M. Using discriminant analysis for multi-class classification: an experimental investigation. Knowl Inf Syst. 2006;10(4):453–72.

[9] Chen B, Jin H, Yang Z, Qu Y, Weng H, Hao T. An approach for transgender population information extraction and summarization from clinical trial text. BMC Med Inf Decis Mak. 2019;19-S(2):159–70.

[10] Tseo Y, Salkola M I, Mohamed A, et al. Information extraction of clinical trial eligibility criteria 2020; arXiv preprint arXiv:2006.07296.

[11] Kang T, Zhang S, Tang Y, et al. EliIE: an open-source information extraction system for clinical trial eligibility criteria. J Am Med Inform Assoc. 2017;24(6):1062–71.

[12] Luo Z, Johnson SB, Lai AM, et al. Extracting temporal constraints from clinical research eligibility criteria using conditional random fields. In: AMIA annual symposium proceedings. Am Med Inform Assoc. 2011;2011:843.

[13] Luo Z, Yetisgen-Yildiz M, Weng C. Dynamic categorization of clinical research eligibility criteria by hierarchical clustering. J Biomed Inform. 2011;44(6):927–35.

[14] Chuan CH. Classifying eligibility criteria in clinical trials using active deep learning. In: 17th IEEE international conference on machine learning and applications (ICMLA). IEEE 2018;305–310.

[15] E. Richard and B. Reddy, "Text Classification for Clinical Trial Operations: Evaluation and Comparison of Natural Language Processing Techniques," Ther. Innov. Regul. Sci., vol. 55, no. 2, pp. 447–453, Mar. 2021, doi: 10.1007/s43441-020-00236-x.

[16] T. Ling et al., "Interpretable machine learning text classification for clinical computed tomography reports – a case study of temporal bone fracture," Comput. Methods Programs Biomed. Update, vol. 3, p. 100104, 2023, doi: 10.1016/j.cmpbup.2023.100104.

[17] L. Jiang, M. Lan, J. D. Menke, et al., "Text classification models for assessing the completeness of randomized controlled trial publications based on CONSORT reporting guidelines," Sci. Rep., vol. 14, p. 21721, 2024, doi: 10.1038/s41598-024-72130-7.

[18] Y. Yang, S. Jayaraj, E. Ludmir, and K. Roberts, "Text Classification of Cancer Clinical Trial Eligibility Criteria," AMIA Annu. Symp. Proc., vol. 2023, pp. 1304–1313, Jan. 2024.

[19] N. Minni and G. Kumaravelan, "Clinical Text Classification in Biomedical Data Analysis: Methods, Resources, Applications, Challenges and Future Perspectives," J. Appl. Bioanal., vol. 11, no. S6, pp. 633–642, 2025, doi: 10.53555/jab.v11si6.2125.

[20] K. Miok, P. Corcoran, and I. Spasić, "The Value of Numbers in Clinical Text Classification," Mach. Learn. Knowl. Extr., vol. 5, no. 3, pp. 746–762, 2023, doi: 10.3390/make5030040.