Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: jeeemi.org; Vol. 8, No. 3, July 2026, pp: 1077-1092 e-ISSN: 2656-8632

RESEARCH ARTICLE OPEN ACCESS

Semantic-Filtered SMOTE-PSO for Breast Cancer Trial Eligibility Classification

Taslim[1,2] and Mumtazimah Mohamad[2,3]

[1] Faculty of Computer Science, Universitas Lancang Kuning, Pekanbaru, Indonesia
[2] Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Terengganu, Malaysia
[3] Artificial Intelligence Research Centre for Islam Sustainability, Universiti Sultan Zainal Abidin, Terengganu, Malaysia

Corresponding author: Taslim. (e-mail: Taslim@unilak.ac.id), **Author(s) Email**: Mumtazimah Mohamad (e-mail: mumtaz@unisza.edu.my)

Abstract **This study addresses breast cancer clinical trial eligibility classification from free-text criteria under severe class imbalance, a condition that biases learning toward the majority class and complicates screening decisions when false positives and false negatives carry different operational costs. The study evaluates whether semantic plausibility control and optimization improve classification performance and screening-oriented error trade-offs under imbalanced conditions. The main contribution of this study is the proposed BEACoN framework, which integrates semantic-filtered augmentation and PSO-guided optimization within a unified screening-oriented eligibility classification setting. Four BioBERT-BiLSTM variants were evaluated using fixed train-validation-test partitions across three random seeds: a baseline model (M1), SMOTE augmentation (M2), SMOTE with cosine filtering (M2.5), and the proposed BEACoN framework (M3). Performance was evaluated using Precision, Recall, F1, AUROC, and AUPRC with pooled multi-seed statistical analysis to improve robustness and reduce single-seed bias. The evaluated augmentation-based configurations achieved pooled F1 scores up to 0.9381 ± 0.0005, AUROC up to 0.9976 ± 0.0001, and AUPRC up to 0.9808 ± 0.0004, indicating improved screening-oriented classification performance relative to the baseline. However, SMOTE with cosine filtering behaved broadly similarly to standard SMOTE under the evaluated embedding setting, indicating that the selected cosine threshold functioned largely as a permissive constraint, although modest seed-dependent prediction differences were still observed. Although BEACoN did not demonstrate statistically significant superiority over SMOTE in aggregate performance, it provided a more balanced false-positive and false-negative trade-off under comparable classification performance. Overall, the findings suggest that plausibility-controlled augmentation may provide practical value for screening-oriented eligibility classification under severe class imbalance.**

Keywords **BioBERT; clinical trial eligibility classification; class imbalance; semantic filtering; SMOTE-PSO**.

I. Introduction

Clinical trials are essential to evidence-based medicine because they provide the basis for evaluating the safety and effectiveness of new interventions before clinical use [1], [2]. This role is especially important in oncology, where delays in recruitment and screening can slow the evaluation of potentially beneficial therapies [3], [4]. Recent studies have shown growing interest in the use of artificial intelligence to support trial recruitment, retention, and eligibility design, particularly in settings where manual screening remains time-consuming and resource intensive [5], [3], [6]. Eligibility criteria are central to this process because they define who may participate in a study and help preserve both patient safety and scientific validity [7], [8], [9]. In practice, however, these criteria are commonly written in free-text form and often include complex clinical terminology [10], [11], temporal conditions, treatment history requirements, and negated statements. Such characteristics make them difficult to process automatically and continue to limit the efficiency of trial screening [12], [13]. Recent work in eligibility criteria parsing, extraction, and query generation has improved the automatic handling of this type of text, but the problem is far from solved because even small linguistic differences can alter clinical meaning and affect downstream decisions [14], [11], [15], [16], [13]. This difficulty becomes more pronounced in screening-oriented classification tasks. In this setting, models must do more than separate two labels. They must distinguish clinically meaningful differences between inclusion and exclusion statements while preserving the context in which those statements appear [6], [17]. Prior studies on eligibility classification and patient-to-

Manuscript received 23 April 2026; Revised 10 June 2026; Accepted 20 June 2026; Available online 6 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1706

trial matching suggest that performance depends not only on representation quality, but also on whether the model can retain medically relevant semantic cues across heterogeneous criteria statements [18], [16], [15]. The consequences of error are also unequal. False negatives may exclude potentially eligible candidates, whereas false positives can increase reviewer workload and reduce screening efficiency [18], [15], [3].

Another issue is class imbalance. In realistic trial datasets, eligible cases are often much less frequent than non-eligible cases [19], [20]. This imbalance can bias a classifier toward the majority class and weaken its ability to detect minority instances [21], [22]. Recent reviews in medical machine learning identify class imbalance as a major factor affecting predictive performance, while recent benchmarking studies show that oversampling can improve results but may also alter classifier behavior in unintended ways [21], [23]. Although oversampling is widely used to address imbalance, standard interpolation-based methods do not explicitly consider the semantic structure of clinical language [24], [25]. When applied to contextual embeddings, they may generate synthetic minority samples that are valid numerically but weak clinically [26], [27]. This limitation matters in eligibility classification because poorly placed synthetic samples can increase class overlap, weaken the effective decision boundary, and raise false positive burden in screening-oriented settings [28], [29]. Despite recent progress in eligibility parsing and extraction, relatively little attention has been given to integrating imbalance-aware augmentation, semantic plausibility control, and optimization within a unified screening-oriented eligibility classification framework [13], [11], [16], [21], [23].

To address this problem, this study proposes BEACoN (BERT-Enhanced Augmentation with Optimization and Semantic Filtering), a screening-oriented augmentation-control framework for breast cancer clinical trial eligibility classification. BEACoN combines biomedical contextual representation with oversampling, cosine-based semantic filtering, and metaheuristic optimization within a unified workflow. The framework is based on the assumption that augmentation quality and class balancing should not be treated as separate decisions, because both can influence the learned decision boundary and the resulting distribution of classification errors. This consideration is particularly important in screening applications, where improvements in aggregate performance may still be undesirable if they substantially increase false-positive burden. Accordingly, BEACoN is evaluated not only through aggregate classification metrics, but also through repeated-run variability and FP/FN redistribution behavior under class imbalance [16], [21], [23]. This study makes four main contributions. First, it compares a BioBERT-BiLSTM baseline with three augmentation settings, namely standard SMOTE, SMOTE with cosine-based semantic filtering, and the full BEACoN framework. Second, it examines variability across repeated runs under a fixed data split. Third, it supplements standard evaluation metrics with uncertainty-aware statistical analysis. Fourth, it analyzes false positives and false negatives to clarify the practical implications of augmentation for trial screening.

Unlike prior studies that evaluate oversampling, semantic filtering, or hyperparameter optimization separately, BEACoN integrates these components within a unified screening-oriented eligibility classification framework. Rather than introducing new individual algorithms, the contribution of this study lies in evaluating optimization-guided augmentation control under clinically imbalanced screening conditions, particularly with respect to false-positive and false-negative trade-offs. The results further indicate that performance differences were driven primarily by augmentation configuration and optimization behavior, while the fixed cosine-based plausibility filter functioned largely as a permissive constraint with limited seed-dependent effects.

The remainder of this paper is organized as follows. Section II describes the dataset, model architecture, augmentation strategies, and evaluation protocol. Section III presents the experimental results and error analysis. Section IV discusses the findings, practical implications, limitations, and future work. Section V concludes the study.

II. Method

A. Dataset and problem formulation

This study addresses binary eligibility classification for breast cancer clinical trial criteria. Each input consists of a short free-text eligibility statement, and the output indicates whether the statement supports eligibility for a breast cancer trial. The dataset was derived from the publicly available labeledEligibilitySample1000000 corpus, which contains one million labeled eligibility statements in its original release. This corpus was compiled from cancer-related interventional trial protocols registered on ClinicalTrials.gov and made available through a public Kaggle mirror [30]. In all experiments, model inputs were restricted to free-text eligibility statements. Structured registry fields were used for disease-level filtering during dataset construction and were not included as predictive features. To reduce statement-level leakage, exact duplicate eligibility statements were removed prior to dataset partitioning.

Manuscript received 23 April 2026; Revised 10 June 2026; Accepted 20 June 2026; Available online 6 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1706

Corpus preparation follows a leakage-control workflow. First, light text normalization is applied, followed by cross-source deduplication using a normalized textual key, yielding 960,424 unique statements. Next, breast cancer records are identified using a precision-first rule-based filter, bc_strict v3b, yielding 75,999 breast cancer-related statements. To resolve cross-subset duplicates, records that appear in both eligible and not eligible subsets are reconciled by retaining the eligible version, producing a breast cancer core of 36,977 eligible and 38,792 not eligible breast cancer statements. To construct a realistic screening setting, the negative class was expanded using non-breast-cancer statements from the deduplicated not-eligible pool. After excluding breast-cancer-related not-eligible instances, 200,000 not-eligible statements from other cancer types were selected, and 60,225 not-eligible non-cancer statements were included based on availability. The final dataset contained 335,994 statements, with Eligible Breast Cancer as the positive class and the composite Not Eligible class as the negative class, yielding a positive-to-negative class ratio of approximately 1:8.1. The dataset was partitioned into training, validation, and test subsets using a stratified 70-15-15 split performed after deduplication to reduce the risk of leakage from exact duplicates. After partitioning, we additionally verified that no exact textual overlap existed across the training, validation, and test sets. Table 1 summarizes the final dataset composition. The train-validation-test partitions were frozen after dataset preparation and reused across all experiments to ensure paired comparison under identical data conditions. Because the final split artifacts retained only statement text and labels, strict protocol-level partitioning using trial identifiers such as NCT IDs could not be enforced and is therefore acknowledged as a limitation.

Table 1. Composition of the final breast cancer eligibility dataset used for classification.

Class	Subtype	Number of Samples
Positive	eligible_bc	36,977
Negative	not_eligible_bc	38,792
Negative	not_eligible_other_cancer	200,000
Negative	not_eligible_non_cancer	60,225
Total		335,994

Fig. 1 presents the overall experimental workflow and comparative model configurations evaluated in this study. The framework begins with dataset preparation and stratified train-validation-test partitioning, followed by four model configurations ranging from the baseline BioBERT-BiLSTM model to augmentation-based variants incorporating SMOTE, cosine-based semantic filtering, and PSO-guided augmentation optimization. M1 uses the original training data, whereas M2, M2.5, and M3 employ augmentation-based training configurations. All configurations were subsequently evaluated using classification, calibration, and statistical analysis metrics under the same fixed evaluation setting.

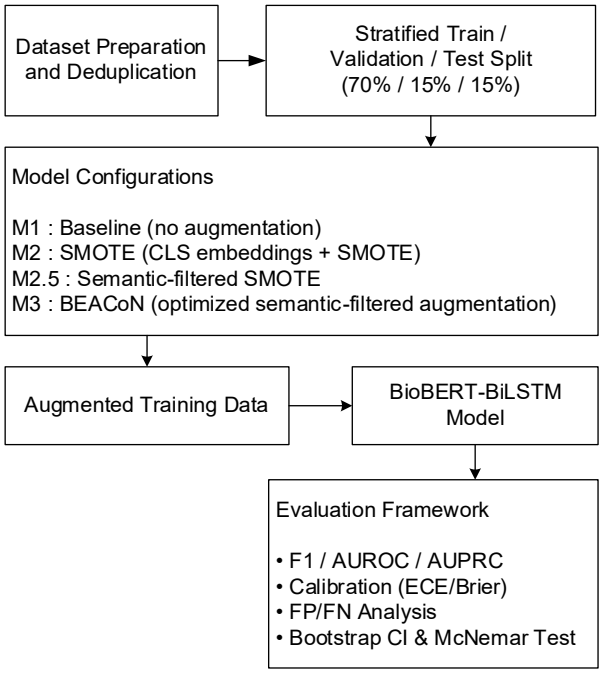

Fig. 1. Overview of the experimental workflow, model configurations, and evaluation framework used in this study.

B. Text preprocessing and representation

Each eligibility criterion is processed as a single statement-level input. During dataset curation, we apply light normalization to reduce superficial variation while preserving clinically meaningful content, including numeric thresholds, medication names, units, and common clinical abbreviations that frequently appear in eligibility text and can influence downstream interpretation [13], [7]. Statements are tokenized using the BioBERT tokenizer and encoded to a fixed maximum sequence length following standard transformer-based preparation for biomedical eligibility text, including truncation, padding, and attention masking [14], [13]. For classification, the downstream model consumes the full sequence of contextual token embeddings produced by BioBERT, that is, the last hidden states. For oversampling and semantic filtering, we derive a fixed-length sentence embedding from the final-layer CLS representation and compute cosine similarity in this embedding space. This choice aligns with established practice in biomedical sentence

Manuscript received 23 April 2026; Revised 10 June 2026; Accepted 20 June 2026; Available online 6 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1706

embedding research, where cosine similarity over dense representations is used to capture semantic relatedness and support downstream similarity-based operations [31]. The CLS representation was selected primarily to maintain computational consistency between augmentation, semantic filtering, and downstream classification stages, rather than as a claim of optimal biomedical sentence embedding quality. However, more specialized sentence-level biomedical representations, including contrastive embedding approaches such as BioSimCSE [31], may provide stronger semantic separation and should be explored in future work.

C. Baseline model architecture (M1)

The baseline classifier combines BioBERT representations with a bidirectional long short-term memory network and a feedforward classification head. The BiLSTM processes token-level embeddings to capture bidirectional dependencies common in eligibility phrasing, including conditional constraints, negation cues, and dispersed clinical qualifiers. The final hidden representation is passed through dropout and a dense layer to produce a single logit. Because the task is formulated as binary classification, the logit is converted into a positive-class probability using the sigmoid function, and the model is optimized using binary cross-entropy with logits. The single-logit output was computed using Eq. (1), following the standard linear formulation commonly used in neural binary classification models [32]. The corresponding positive-class probability was obtained through the sigmoid activation function in Eq. (2) [32], while the final binary prediction was determined using the threshold rule in Eq. (3), following standard binary decision thresholding in neural classification models [32]. Model optimization employed the binary cross-entropy loss defined in Eq. (4), a standard objective function for deep neural binary classification [33].

$$z_i = W_o h_i + b_o \tag{1}$$

$$\hat{p}_i = \sigma(z_i) = \frac{1}{1+\exp(-z_i)} \tag{2}$$

$$\hat{y}_i = 1 \text{ if } \hat{p}_i \geq 0.5, \text{ otherwise } 0 \tag{3}$$

$$L_{BCE} = -\text{mean}[y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)] \tag{4}$$

where z_i is the output logit, h_i is the final hidden representation, W_o and b_o are trainable output-layer parameters, $\hat{p}_i$ is the predicted positive-class probability, $\hat{y}_i$ denotes the predicted binary class label, and y_i is the true binary label.

D. Imbalance handling via oversampling (M2)

To mitigate class imbalance during training, we applied the Synthetic Minority Over-sampling Technique (SMOTE) to increase the effective density of minority-class training instances. SMOTE synthesizes new minority samples by interpolating between a minority instance and its nearest minority neighbors, thereby expanding coverage of the minority region without simple replication [34]. In this study, oversampling was performed in the BioBERT CLS embedding space because eligibility statements often exhibit substantial surface variation, and proximity in contextual embedding space is more likely to reflect semantic relatedness than interpolation over sparse lexical features [31], [17]. The minority embedding set is denoted as $E^+ = \{e_i^+\}_{i=1}^{N_+}$, where e_i^+ represents the BioBERT CLS embedding of the i-th positive-class training statement. In SMOTE, a synthetic minority embedding is generated by interpolating between a minority embedding and one of its nearest minority neighbors, as shown in Eq. (5) [34].

$$\tilde{e}_i = e_i^+ + \lambda\left(e_j^+ - e_i^+\right), \lambda \in [0,1] \tag{5}$$

Because the downstream BioBERT-BiLSTM classifier was trained on textual inputs rather than synthetic vectors, each generated synthetic embedding was mapped back to the nearest original minority-class statement using cosine distance, as shown in Eq. (6) [31].

$$x_i^{map} = \arg\min_{x_j^+ \in X^+} d_{cos}\left(\tilde{e}_i, e_j^+\right) \tag{6}$$

where $\tilde{e}_i$ is the synthetic minority embedding, e_i^+ is the selected minority embedding, e_j^+ is one of its nearest minority neighbors, $\lambda$is a random interpolation coefficient, X^+ is the set of original positive-class training statements, and x_i^{map} is the nearest textual statement used to construct the oversampled training set.

Although SMOTE can improve sensitivity to minority cases, interpolation in embedding space can still introduce semantically weak samples, particularly in clinical eligibility text where labels may hinge on negation, temporal qualifiers, or multi-constraint phrasing. Such drift can increase overlap near decision boundaries and may raise false-positive burden in screening-oriented settings [34], [35]. These limitations motivate the plausibility control and optimization extensions evaluated in subsequent configurations. Because the downstream classifier operated on textual eligibility statements, the mapping stage was designed to preserve valid clinical text inputs while allowing augmentation decisions to be guided by embedding-space interpolation. Consequently, the proposed augmentation strategy should be interpreted as representation-guided augmentation rather than generation of entirely novel synthetic eligibility text.

E. Cosine-based semantic filtering (M2.5)

Manuscript received 23 April 2026; Revised 10 June 2026; Accepted 20 June 2026; Available online 6 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1706

The M2.5 configuration extends SMOTE by applying a cosine-based semantic filtering step in the BioBERT CLS embedding space. After synthetic minority embeddings were generated, each synthetic vector was compared with the original minority embedding set using cosine similarity. The cosine similarity between a synthetic embedding $\tilde{e}_i$ and an original minority embedding e_j^+ is defined in Eq. (7) [31].

$$\cos\left(\tilde{e_i}, e_j^+\right) = \frac{\tilde{e_i} \cdot e_j^+}{\|\tilde{e_i}\| \|e_j^+\|} \quad (7)$$

For each synthetic embedding, the top-m nearest minority neighbors were retrieved based on cosine similarity. The semantic plausibility score was then calculated as the mean similarity between the synthetic embedding and its top-m minority neighbors, as shown in Eq. (8) [31].

$$s_i = \frac{1}{m} \sum_{j=1}^{m} \cos\left(\tilde{e_i}, e_j^+\right) \quad (8)$$

A synthetic embedding was retained only when its mean top-m cosine similarity reached the predefined threshold τ. The binary acceptance indicator is expressed in Eq. (9) [35].

$$a_i = 1 \text{ if } s_i \geq \tau, \text{ otherwise } 0 \quad (9)$$

The acceptance rate of the cosine-based filter was computed as the proportion of generated synthetic embeddings that satisfied the semantic plausibility criterion, as shown in Eq. (10) [35].

$$R_{acc}CC = \frac{1}{N_{syn}} \sum_{i=1}^{N_{syn}} a_i \quad (10)$$

where $\tilde{e}_i$ is the i-th synthetic minority embedding, e_j^+ is an original minority embedding, e_{ij}^+ denotes the j-th nearest minority neighbor of $\tilde{e}_i$, m denotes the number of nearest minority neighbors used in the cosine-based semantic plausibility calculation, s_i is the mean top-m cosine similarity score, a_i is the binary acceptance indicator, τ is the semantic plausibility threshold, and N_{syn} is the number of generated synthetic embeddings.

In this study, the top-neighbor parameter m was set to 5 and the cosine threshold τ was set to 0.85. Synthetic embeddings that satisfied the filtering criterion were mapped back to valid textual inputs by selecting the nearest original minority statement associated with the closest minority embedding. This mapping ensured that the downstream BioBERT-BiLSTM classifier remained trained on textual eligibility statements rather than on synthetic embedding vectors. M2.5 was included as an ablation setting to isolate the effect of cosine-based plausibility filtering from PSO-guided optimization. Under the evaluated configuration, the filter was largely non-restrictive: all 183,427 generated synthetic embeddings satisfied the filtering criterion, yielding an acceptance rate of 1.0. The minimum observed mean top-5 cosine similarity was 0.942, indicating that the selected threshold did not exclude any synthetic embeddings under the current CLS representation and mapping procedure. The observed behavior indicates that the selected threshold operated within a permissive region of the embedding space, such that stronger cosine constraints would likely be required before semantic filtering meaningfully alters the augmentation set. However, permissive augmentation acceptance does not necessarily imply identical downstream classifier behavior across all random seeds, because stochastic training dynamics may still produce modest prediction-level differences.

F. BEACoN framework with optimization and semantic filtering (M3)

BEACoN, short for BERT-Enhanced Augmentation with Optimization and Semantic Filtering, is the complete augmentation framework proposed in this study. The framework integrates SMOTE-style minority oversampling in the BioBERT CLS embedding space, cosine-based semantic plausibility filtering, and Particle Swarm Optimization (PSO) to tune the augmentation configuration [35], [36], [37]. In contrast to M2.5, which applies a fixed filtering threshold, BEACoN jointly optimizes the balancing strength and plausibility constraint so that oversampling is guided by validation performance rather than by a manually fixed configuration. The PSO search space is defined by three augmentation hyperparameters: the SMOTE neighbor parameter k_{smote}, the sampling ratio r, and the semantic filtering threshold τ_{syn}. The PSO candidate solution vector formulation is expressed in Eq. (11), following the standard particle swarm optimization representation proposed in prior optimization studies [36].

$$\theta_i = \left(k_{smote,i}, r_i, \tau_{syn,i}\right) \quad (11)$$

In this study, the search bounds were defined as $k_{\text{smote}} \in [3,15]$, $r \in [0.25,1.00]$, and $\tau_{\text{syn}} \in [0.70,0.95]$. During optimization, each particle updated its velocity based on the current velocity, personal best position, and global best position, as shown in Eq. (12) [36], [37].

$$v_i^{t+1} = \omega v_i^t + c_1 r_1 \left(p_i^t - \theta_i^t\right) + c_2 r_2 \left(g^t - \theta_i^t\right) \quad (12)$$

The particle position was then updated using Eq. (13) [36], [37].

$$\theta_i^{t+1} = \theta_i^t + v_i^{t+1} \quad (13)$$

where θ_i^t represents the position of the i-th particle at iteration t, v_i^t denotes the particle velocity, p_i^t indicates

Manuscript received 23 April 2026; Revised 10 June 2026; Accepted 20 June 2026; Available online 6 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1706

the personal best solution obtained by the i-th particle, and g^t corresponds to the global best solution identified by the swarm. Furthermore, ω is the inertia coefficient controlling momentum preservation, while c_1 and c_2 denote the cognitive and social acceleration coefficients, respectively. The variables r_1 and r_2 are uniformly distributed random values sampled from $U(0,1)$. To reduce computational cost, candidate configurations were first evaluated using a CLS-based surrogate classifier. For each candidate θᵢ, SMOTE and cosine filtering were applied according to the candidate parameters, and the resulting augmented BioBERT CLS embedding representation was evaluated on the validation set using a Logistic Regression surrogate classifier. The optimization fitness was defined as validation AUPRC because AUPRC is informative for assessing minority-positive ranking quality under class imbalance, as shown in Eq. (14) [32]. The surrogate stage was intended to provide a computationally efficient proxy for ranking augmentation configurations during PSO search rather than to replace the final BioBERT-BiLSTM classifier. After optimization, the selected augmentation configuration was used to retrain and evaluate the full downstream BioBERT-BiLSTM model.

$$F(\theta_i) = AUPRC_{val}(\theta_i) \tag{14}$$

The optimal augmentation configuration was selected as the candidate that maximized validation AUPRC, as shown in Eq. (15) [32].

$$\theta^* = \arg\max_{\theta_i} F(\theta_i) \tag{15}$$

where $F(\theta_i)$ is the validation fitness of candidate θ_i, and θ^* is the selected BEACoN configuration. After PSO identified θ^*, the final BioBERT-BiLSTM model was retrained using the optimized augmentation setting. Thus, the surrogate stage was used only for configuration screening, whereas the final reported results were produced by the same BioBERT-BiLSTM evaluation protocol used for the other model variants. To improve methodological reproducibility and provide a clearer overview of the operational sequence underlying the proposed augmentation strategy, Algorithm 1 summarizes the augmentation, semantic filtering, surrogate-based evaluation, and optimization workflow associated with the BEACoN configuration within the broader experimental pipeline used throughout this study. The pseudocode was designed to highlight the interaction between augmentation control, filtering consistency, and optimization-guided configuration selection across the evaluated experimental settings.

Algorithm 1. BEACoN augmentation and optimization workflow.

Input: D_train, D_val, BioBERT encoder
Output: Optimized augmentation configuration theta_star

Extract CLS embeddings from minority-class statements
Initialize PSO particles theta_i = (k_i, r_i, τ_i)
for each PSO iteration:
    Generate SMOTE embeddings using theta_i
    Apply cosine-based semantic filtering
    Map retained embeddings to nearest minority statements
    Construct augmented training set
    Evaluate validation AUPRC using LR surrogate classifier
    Update particle fitness and global best solution
end for
Select optimal configuration theta_star
Retrain BioBERT-BiLSTM using optimized setting
Evaluate final model on the fixed test set

G. Experimental protocol and reproducibility

All model variants were evaluated on the same fixed train-validation-test split to enable paired comparison on an identical test set. To assess variability across random initialization, each configuration was trained and evaluated using three random seeds, 42, 100, and 2024. For each seed, library and framework random states were initialized before training to support reproducibility.

The downstream classifier used the BioBERT checkpoint dmis-lab/biobert-base-cased-v1.1 with a maximum sequence length of 128 tokens. The contextual token embeddings produced by BioBERT were processed using a single-layer BiLSTM with a hidden size of 128, followed by a dropout layer with dropout probability 0.2 and a dense output layer for binary classification. Model optimization was performed using AdamW with a learning rate of 2×10^{-5}, batch size 32, and maximum training length of 5 epochs. Early stopping with patience 2 based on validation F1 score was applied to reduce overfitting. For PSO-based augmentation optimization, the swarm size and maximum iteration count were both set to 10. The inertia coefficient ω was set to 0.7, while the cognitive and social acceleration coefficients c1 and c2 were both set to 1.4. To reduce computational cost during configuration search, candidate augmentation settings were first evaluated using a Logistic Regression surrogate classifier trained on augmented BioBERT CLS embedding representations before retraining the final BioBERT-BiLSTM model using the selected augmentation configuration. The surrogate

Manuscript received 23 April 2026; Revised 10 June 2026; Accepted 20 June 2026; Available online 6 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1706

classifier was used only for computationally efficient configuration screening during PSO optimization, whereas all final reported performance results were obtained using the full BioBERT-BiLSTM evaluation pipeline. All experiments were conducted in a Google Colab environment using NVIDIA Tesla T4 GPU acceleration with CUDA 13.0 support and approximately 15 GB GPU memory. The implementation environment used Python 3.12.13, PyTorch 2.10.0, Transformers 5.0.0, Scikit-learn 1.6.1, and Imbalanced-learn 0.14.1.

H. Evaluation metrics and statistical analysis

Performance was evaluated using Precision, Recall, F1 score, Accuracy, AUROC, and AUPRC. These metrics were selected to provide a complementary assessment of classifier behavior under class imbalance, where no single metric is sufficient to capture both minority-class detection and false-positive burden. In this setting, Precision, Recall, and F1 score describe threshold-dependent classification behavior, while AUROC and AUPRC summarize discrimination across decision thresholds. Recent studies have emphasized the importance of using multiple evaluation measures when assessing predictive models on imbalanced data, particularly in medical and screening-related applications [39]. Precision quantifies the proportion of predicted positive statements that are truly positive, as shown in Eq. (16) [39].

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

Recall quantifies the proportion of true positive statements that are correctly identified by the classifier, as shown in Eq. (17) [39].

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

The F1 score summarizes the balance between Precision and Recall, as shown in Eq. (18) [39].

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{18}$$

Accuracy was also reported as a general measure of the proportion of correctly classified instances, as shown in Eq. (19) [39].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{19}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. In addition to threshold-dependent metrics, AUPRC was used to summarize precision-recall behavior across decision thresholds. AUPRC is particularly informative under class imbalance because it focuses on positive-class retrieval quality rather than overall correctness alone [39]. The area under the precision-recall curve can be expressed as the integral of precision over recall, as shown in Eq. (20), following standard AUPRC formulation in imbalanced classification evaluation [39].

$$AUPRC = \int_0^1 Precision(Recall)\ dRecall \tag{20}$$

where $Precision(Recall)$ denotes precision as a function of recall along the precision-recall curve.

Because all models were evaluated on the same fixed test split and trained across multiple random seeds, uncertainty was quantified using pooled bootstrap 95 percent confidence intervals computed from resampled predictions across runs. This approach was used to complement point estimates and to reflect variability arising from both test-set composition and repeated model training. Reporting uncertainty alongside aggregate metrics is increasingly recommended in predictive modeling studies, especially in clinical contexts where small numerical differences may not correspond to stable performance differences [33], [38]. For a metric value M, the percentile-based bootstrap confidence interval was computed from the empirical bootstrap distribution of resampled estimates, as shown in Eq. (21), following standard bootstrap interval estimation procedures [38].

$$CI_{95\%}(M) = [Q_{0.025}(M^*), Q_{0.975}(M^*)] \tag{21}$$

where M^* is the bootstrap distribution of the evaluated metric, while $Q_{0.025}$ and $Q_{0.975}$ are the lower and upper percentile bounds of the 95% confidence interval, respectively. To assess whether observed performance differences corresponded to consistent decision changes on the same test instances, paired comparisons between model outputs were conducted using McNemar tests. This test is appropriate when competing classifiers are evaluated on identical samples because it focuses on discordant prediction pairs rather than independent aggregate scores [39]. The McNemar test statistic used for paired prediction comparison on identical test instances is shown in Eq. (22) [39].

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}} \tag{22}$$

where n_{01} is the number of instances misclassified by the first model but correctly classified by the second model, and n_{10} is the number of instances correctly classified by the first model but misclassified by the second model. Statistical testing was reported together with aggregate metrics to support a more robust interpretation of model performance in screening-oriented settings [39], [32].

III. Result

The four model configurations were evaluated on the same fixed test split across three random seeds,

Manuscript received 23 April 2026; Revised 10 June 2026; Accepted 20 June 2026; Available online 6 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1706

namely 42, 100, and 2024. The evaluated models consisted of the BioBERT-BiLSTM baseline (M1), SMOTE augmentation (M2), SMOTE with cosine-based filtering (M2.5), and the proposed BEACoN framework (M3). Across the three seeds, the augmentation-based models mainly improved F1 performance and altered the balance between false positives and false negatives, while AUROC and AUPRC remained consistently high across all configurations.

Table 2 summarizes the aggregate performance across three random seeds. SMOTE improved the baseline F1 from 0.9313 ± 0.0024 to 0.9365 ± 0.0013, while the cosine-filtered SMOTE configuration achieved a slightly higher F1 of 0.9381 ± 0.0005. BEACoN achieved the highest Accuracy, 0.9862 ± 0.0005, with a comparable F1 score of 0.9379 ± 0.0029. AUROC and AUPRC remained consistently high across all configurations, ranging from 0.9974 to 0.9976 and from 0.9759 to 0.9808, respectively. These results suggest that the main distinction among the augmentation-based configurations lies less in large changes to threshold-independent ranking performance and more in how false-positive and false-negative trade-offs were redistributed under comparable aggregate metrics.

Table 2. Overall classification performance across seeds for all evaluated model configurations.

Model	Accuracy	F1	AUROC	AUPRC
M1	0.9849 ± 0.0004	0.9313 ± 0.0024	0.9975 ± 0.0002	0.9759 ± 0.0051
M2	0.9858 ± 0.0002	0.9365 ± 0.0013	0.9974 ± 0.0003	0.9773 ± 0.0050
M2.5	0.9861 ± 0.0001	0.9381 ± 0.0005	0.9976 ± 0.0001	0.9808 ± 0.0004
M3	0.9862 ± 0.0005	0.9379 ± 0.0029	0.9974 ± 0.0002	0.9804 ± 0.0021

Table 3. Pooled bootstrap 95% confidence intervals for all evaluation metrics.

Metric	M1	M2	M2.5	M3
F1	0.9287–0.9341	0.9338–0.9392	0.9354–0.9408	0.9352–0.9405
AUROC	0.9968–0.9972	0.9967–0.9972	0.9972–0.9976	0.9971–0.9974
AUPRC	0.9728–0.9766	0.9743–0.9780	0.9772–0.9808	0.9764–0.9801

The pooled bootstrap confidence intervals in Table 3 show that augmentation produced consistent improvements in F1 relative to the baseline configuration, while AUROC and AUPRC remained uniformly high across all models. The baseline F1 interval was 0.9287–0.9341, compared with 0.9338–0.9392 for SMOTE, 0.9354–0.9408 for the cosine-filtered SMOTE configuration, and 0.9352–0.9405 for BEACoN. The substantial overlap among the augmentation-based intervals suggests that the primary differences among these configurations lie less in large changes to threshold-independent ranking performance and more in how false-positive and false-negative trade-offs were redistributed.

Table 4 presents pooled McNemar significance tests across the evaluated configurations. All augmentation-based models demonstrated statistically significant prediction differences relative to the baseline configuration. The strongest difference relative to the baseline was observed for the cosine-filtered SMOTE configuration (M2.5), consistent with its more recall-oriented operating-point behavior. In contrast, pairwise comparisons among the augmentation-based configurations were not statistically significant, suggesting that the principal distinctions among these models arise less from large changes in aggregate classification accuracy and more from differences in false-positive and false-negative redistribution.

Table 4. Pooled McNemar test results for paired model comparisons across seeds.

Comparison	n01	n10	Test statistic	p-value
M1 vs M2	796	935	11.0017	0.000910
M1 vs M2.5	817	996	17.4760	0.000029
M1 vs M3	727	923	23.0455	0.000002
M2 vs M3	760	817	1.9886	0.158489
M2.5 vs M3	782	799	0.1619	0.687392

Additional comparison between M2 and M2.5 showed highly similar overall prediction behavior, with no statistically significant paired prediction differences observed in the pooled evaluation ($p = 0.1161$). These findings support the interpretation that the fixed cosine threshold operated largely as a permissive constraint under the evaluated embedding setting, although limited seed-dependent prediction-level differences were still observed. Table 5 summarizes the false-positive and false-negative distributions across all evaluated seeds, while Fig. 2 visualizes the representative Seed 42 FP/FN behavior. Standard SMOTE (M2) generally reduced false negatives relative to the baseline, although the magnitude of the FP/FN redistribution varied across seeds. Under Seed 42, the cosine-filtered configuration (M2.5) produced the strongest reduction in false negatives (200) but also the largest false-positive increase (500). BEACoN moderated this imbalance by limiting false-positive growth while maintaining lower false-negative counts than the baseline across evaluated runs. These results suggest that screening-oriented evaluation should consider not only aggregate performance, but also how augmentation changes the balance between false positives and false negatives in practical workflows. Precision–Recall analysis further clarified the

Manuscript received 23 April 2026; Revised 10 June 2026; Accepted 20 June 2026; Available online 6 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1706

operational differences among the evaluated augmentation strategies. Standard SMOTE (M2) improved minority-positive sensitivity primarily by reducing false negatives, indicating a moderate shift toward recall-oriented screening behavior. Under Seed 42, the cosine-filtered configuration (M2.5) emphasized recall more aggressively, further lowering false negatives at the cost of increased false-positive burden. In contrast, BEACoN retained much of the sensitivity improvement while moderating false-positive expansion, resulting in a more balanced screening profile across the evaluated runs. These findings indicate that augmentation strategies influence not only aggregate discrimination performance, but also the redistribution of clinically relevant prediction errors. Collectively, the results suggest that practical screening utility depends on balancing missed eligible cases against unnecessary manual review workload rather than optimizing global performance metrics alone.

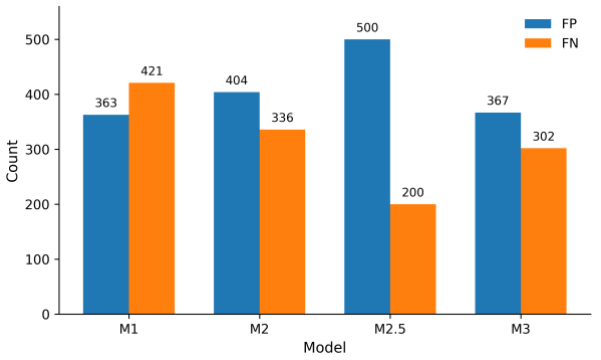

Fig. 2. False-positive and false-negative redistribution patterns under Seed 42 across evaluated model configurations.

Table 5. False-positive and false-negative distributions across evaluated seeds for all model configurations.

Model	Seed	FP	FN
M1	42	363	421
M1	100	337	419
M1	2024	404	336
M2	42	404	336
M2	100	275	422
M2	2024	252	452
M2.5	42	500	200
M2.5	100	275	422
M2.5	2024	252	452
M3	42	367	302
M3	100	344	380
M3	2024	463	228

Although M2 and M2.5 produced identical FP/FN counts for Seeds 100 and 2024, prediction-level analysis still showed limited stochastic divergence under Seed 42, indicating that permissive semantic filtering did not fully eliminate downstream variability. To further assess probability reliability under screening-oriented eligibility classification, calibration performance was evaluated using pooled predictions across the three random seeds (42, 100, and 2024). Calibration quality was assessed using Brier Score and Expected Calibration Error (ECE), with lower values indicating better alignment between predicted probabilities and observed outcomes.

Table 6 summarizes the pooled calibration results. All models exhibited low Brier Score and ECE values, indicating reasonably well-calibrated probability estimates. While the baseline model (M1) achieved the lowest ECE value, BEACoN (M3) produced the lowest Brier Score while preserving the screening-oriented FP/FN trade-off advantages observed previously. These findings indicate that augmentation-based configurations did not substantially degrade probability reliability under severe class imbalance conditions. Calibration behavior was also broadly consistent with the aggregate classification results. M2 and M2.5 demonstrated highly similar calibration characteristics, suggesting that the selected cosine threshold operated largely as a permissive constraint under the evaluated embedding setting, although Seed 42 prediction-level differences indicate that minor stochastic variations may still emerge despite comparable aggregate calibration behavior.

Fig. 3 further illustrates the pooled reliability behavior across the evaluated configurations, showing that the predicted probabilities remained broadly aligned with observed outcomes despite differences in augmentation strategy.

IV. Discussion

A. Interpretation of findings

The results show that imbalance-aware augmentation improved F1 under the evaluated breast cancer eligibility screening setting. The baseline achieved an F1 score of 0.9313 ± 0.0024, compared with 0.9365 ± 0.0013 for SMOTE, 0.9381 ± 0.0005 for the cosine-filtered SMOTE configuration, and 0.9379 ± 0.0029 for BEACoN. The pooled bootstrap intervals showed the same pattern, supporting the view that class imbalance remains important for minority-class detection in medical machine learning [40]. The paired statistical analysis suggests that augmentation altered prediction behavior beyond small numerical differences in aggregate metrics. Pooled McNemar tests showed significant differences between the baseline and all augmentation-based configurations, including M1 vs M2 (test statistic = 11.0017, p = 0.000910), M1 vs M2.5 (test statistic = 17.4760, p = 0.000029), and M1 vs M3

Manuscript received 23 April 2026; Revised 10 June 2026; Accepted 20 June 2026; Available online 6 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1706

(test statistic = 23.0455, p = 0.000002), whereas the comparison between SMOTE and BEACoN was not statistically significant (M2 vs M3: test statistic = 1.9886, p = 0.158489). These findings suggest that BEACoN primarily improved screening behavior through a more balanced redistribution of false positives and false negatives under comparable aggregate performance.

The error-level results further clarify this distinction.

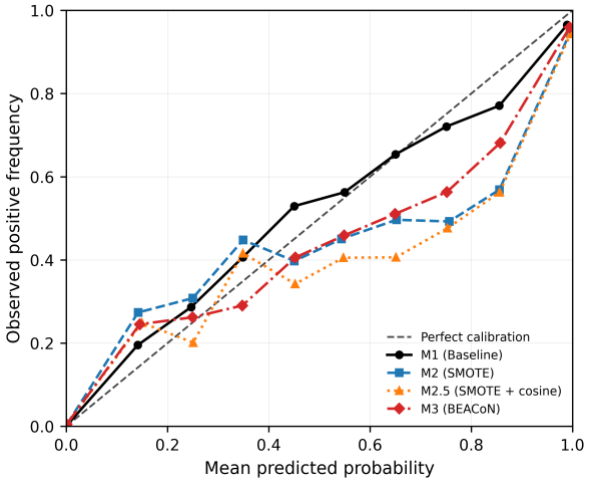

Fig. 3. Reliability diagram comparing pooled calibration behavior across all evaluated models using predictions aggregated across the three random seeds.

Table 6. Calibration performance across evaluated model configurations.

Model	Brier Score	ECE
M1	0.011840	0.004895
M2	0.012172	0.009914
M2.5	0.012060	0.010677
M3	0.011531	0.008699

Under Seed 42, the baseline produced 363 false positives and 421 false negatives. Standard SMOTE (M2) reduced false negatives to 336 while moderately increasing false positives to 404. The cosine-filtered configuration (M2.5) further reduced false negatives to 200 but produced substantially higher false-positive counts (500). By comparison, BEACoN produced a more balanced error profile, with 367 false positives and 302 false negatives, compared with 500 false positives and 200 false negatives under M2.5. Compared with M2.5, BEACoN substantially reduced false-positive burden while preserving much of the recall improvement relative to the baseline. This distinction matters operationally because missed eligible statements and unnecessary downstream review carry different screening costs.

Although Fig. 2 presents Seed 42 as a representative FP/FN visualization, Table 5 reports the FP/FN distributions across all evaluated seeds. Together with the pooled F1 estimates, confidence intervals, and McNemar analyses, these results indicate that the augmentation-based configurations remained broadly consistent in aggregate behavior despite limited seed-dependent prediction differences. Thus, the SMOTE–BEACoN comparison is better interpreted as a difference in screening behavior rather than as a simple ranking of aggregate performance. The cosine-filtered configuration produced the strongest sensitivity-oriented behavior under Seed 42, with the lowest false-negative count but the highest false-positive burden. BEACoN preserved part of that sensitivity gain while keeping false positives substantially closer to the baseline level. For workflows with limited reviewer capacity, this more balanced error profile may be preferable over a marginal difference in aggregate metrics [41]. Residual classification errors remained particularly associated with negation patterns, temporal constraints, abbreviated biomarker terminology, and multi-condition clinical requirements. These characteristics can produce semantically overlapping embeddings despite clinically different eligibility interpretations. Collectively, these findings suggest that screening-oriented eligibility classification depends not only on imbalance handling, but also on preserving fine-grained clinical context across heterogeneous eligibility phrasing.

The comparison between standard SMOTE and cosine-filtered SMOTE provides an informative ablation result. Under the evaluated BioBERT CLS embedding setting, cosine filtering functioned largely as a permissive plausibility constraint, producing only limited seed-dependent prediction differences. Although Seed 42 exhibited a more recall-oriented operating-point shift, pooled F1 estimates, calibration behavior, and paired statistical analyses remained broadly similar across the two configurations, indicating limited downstream influence of the evaluated cosine threshold. Consequently, the observed differences between SMOTE and BEACoN are better interpreted as effects of optimization-guided augmentation control rather than filtering selectivity alone. Under the current embedding representation, substantially stronger plausibility constraints would likely be required for semantic filtering to meaningfully alter the augmentation set. Additional inspection of the nearest-neighbor reconstruction outputs showed that the final augmented training set contained 261,915 samples mapped to 235,195 unique statements. Despite the embedding-based reconstruction process, the augmentation pipeline retained substantial textual diversity and did not collapse into repeated duplication of minority samples. Collectively, these observations

Manuscript received 23 April 2026; Revised 10 June 2026; Accepted 20 June 2026; Available online 6 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1706

support interpreting the evaluated augmentation configuration as a representation-guided expansion strategy rather than a simple duplication mechanism.

B. Comparison with prior studies

The findings are consistent with recent work showing that eligibility criteria text remains difficult to model because clinical meaning often depends on interacting constraints, including negation, temporal qualifiers, and nested conditions [16]. The consistently high AUROC values, ranging from 0.9974 to 0.9976, and AUPRC values, ranging from 0.9759 to 0.9808, indicate that BioBERT-BiLSTM provided strong ranking performance for this task. However, the F1 improvement after augmentation shows that representation quality alone does not fully address the effect of class imbalance. This supports prior evidence that domain-adapted contextual encoders are useful for eligibility criteria processing, but they still require careful handling of downstream classification conditions when positive cases are sparse [14]. Related patient-trial matching studies have emphasized that eligibility assessment is not only a text classification problem, but also a screening task in which different error types carry distinct practical costs [18], [10]. Consistent with this perspective, the present findings indicate that augmentation strategies should be evaluated not only by sensitivity improvement, but also by their effect on downstream review burden. In this setting, BEACoN moderated the false-positive increase associated with SMOTE while maintaining improved sensitivity relative to the baseline.

Recent transformer-based and LLM-based approaches to eligibility-related NLP have likewise highlighted the importance of semantic alignment between free-text criteria and model decisions [42], [15]. The present results further show that the fixed cosine-filtered M2.5 configuration behaved broadly similarly to standard SMOTE, with pooled F1 scores of 0.9381 ± 0.0005 and 0.9365 ± 0.0013, respectively, whereas BEACoN extended this setting through PSO-guided augmentation selection based on validation AUPRC. Recent studies in the imbalance-learning literature have shown that oversampling can improve minority-class sensitivity while simultaneously increasing false positives through decision-boundary expansion [43]. The present findings are consistent with that observation. Under Seed 42, standard SMOTE (M2) moderately reduced false negatives while maintaining a false-positive burden close to the baseline, whereas the cosine-filtered configuration (M2.5) produced a more recall-oriented operating point that further reduced false negatives at the cost of increased false positives. Relative to M2.5, BEACoN reduced false positives from 500 to 367 while preserving substantially fewer false negatives than the baseline. Compared with M2, BEACoN maintained a comparable false-positive burden while further reducing false negatives. Overall, these findings support evaluating augmentation quality not only through aggregate metrics, but also through the distribution of clinically relevant errors.

Table 7 summarizes representative prior studies related to eligibility-focused biomedical NLP and clinical trial screening.

Table 7. Comparison of the current study with representative prior studies.

Study	Focus	Aug.-aware	Screening-oriented
Bornet et al. [17]	Clustering	✕	✕
Chen et al. [44]	Trial matching	✕	Partial
Han et al. [16]	Eligibility classification	✕	✕
Lu et al. [45]	Imbalance NLP	Partial	✕
Peluso et al. [46]	Calibration-aware NLP	✕	✓
Current study	Augmentation control	✓	✓

As summarized in Table 7, prior studies have demonstrated the value of biomedical transformer and semantic representation approaches for eligibility-related NLP tasks, including semantic clustering, trial matching, eligibility classification, and calibration-aware evaluation [16], [17], [44], [46]. Consistent with these findings, the evaluated augmentation-based configurations maintained strong AUROC and AUPRC performance while improving minority-class detection under severe class imbalance. However, the present study additionally examined screening-oriented false-positive and false-negative trade-offs, calibration-oriented evaluation, and optimization-guided augmentation within a unified breast cancer eligibility classification framework. These findings suggest that threshold-independent ranking metrics alone may be insufficient for screening-oriented eligibility assessment, where false-positive and false-negative distributions also influence practical screening utility. Dataset construction also affects the interpretation of augmentation results. Exact duplicate statements were removed before dataset partitioning, and the final train-validation-test splits were verified to contain no exact textual overlap. This is important because duplicate and near-duplicate registry statements can inflate downstream performance estimates if not controlled before screening [47]. Under this leakage-controlled setting, the observed differences among the baseline, SMOTE, M2.5, and BEACoN are therefore more likely

Manuscript received 23 April 2026; Revised 10 June 2026; Accepted 20 June 2026; Available online 6 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1706

to reflect imbalance-related model behavior rather than repeated statements across data splits. However, because the final split artifacts retained only statement text and labels, protocol-level leakage using trial identifiers could not be fully excluded. The present results should therefore be interpreted as leakage-controlled at the exact statement level rather than at the strict protocol level.

Operationally, the choice among M2, M2.5, and BEACoN depends on screening priorities. Standard SMOTE (M2) provided a moderate reduction in false negatives with a relatively limited increase in false positives, whereas M2.5 produced a more aggressive recall-oriented operating point with substantially higher false-positive burden. BEACoN may therefore be preferable in screening workflows where reviewer workload is a major constraint, because it preserved much of the recall improvement while moderating false-positive growth. The present results support interpreting BEACoN as a quality-aware augmentation strategy rather than as a method designed only to maximize aggregate scores. In practice, such models may be more suitable as prioritization or triage support tools that assist reviewer workflow rather than as fully autonomous eligibility decision systems. A deployment strategy could therefore define an acceptable recall or false-negative target first, then select the configuration that minimizes false positives under that constraint.

Several limitations should be considered. First, the experiments were conducted on a disease-focused corpus of eligibility statements derived from ClinicalTrials.gov-related data and reflected statement-level eligibility screening rather than end-to-end patient-trial matching. Consequently, the findings do not directly establish performance on patient-level or multi-source clinical data. Second, semantic plausibility control was based on BioBERT CLS embeddings and cosine similarity, which may not fully capture fine-grained clinical distinctions in eligibility statements containing multiple constraints. Under the evaluated embedding setting, the cosine-filtered M2.5 configuration behaved similarly to standard SMOTE, suggesting that the evaluated cosine threshold functioned largely as a permissive constraint within the current representation space. Third, although final test results were obtained using the full BioBERT-BiLSTM model, BEACoN employed a surrogate evaluation stage to reduce the computational cost of PSO-based configuration search. In addition, the study did not include external validation across independent datasets or broader clinical settings. Therefore, the findings should be interpreted within a leakage-controlled single-corpus evaluation setting rather than as evidence of broad deployment-level generalization.

Finally, because synthetic embeddings were mapped back to nearest minority statements, the augmentation process may increase the reuse of semantically similar minority samples rather than generating fully novel textual examples. Consequently, the introduced augmentation may provide limited embedding-space diversity under the evaluated representation setting.

Future work should examine utility-based optimization objectives that more directly reflect screening priorities under severe class imbalance. For example, optimization strategies could incorporate asymmetric false-positive and false-negative costs or recall-constrained objectives to better align model selection with screening workload and sensitivity requirements. Further investigation is also needed for more selective plausibility controls. In the present study, the fixed cosine threshold used in M2.5 behaved similarly to a permissive constraint and produced only limited downstream differences relative to standard SMOTE. Future approaches may benefit from combining positive-neighborhood similarity with negative-neighborhood separation to better reduce boundary overlap and false-positive burden. Finally, additional validation across independent eligibility datasets, temporal splits, protocol-level partitioning, and broader disease settings will be important for assessing generalization and deployment-level robustness in screening-oriented eligibility classification systems.

V. Conclusion

This study evaluated imbalance-aware augmentation strategies for breast cancer clinical trial eligibility classification using fixed-split repeated multi-seed experiments. Augmentation-based configurations consistently improved minority-class detection relative to the BioBERT-BiLSTM baseline, achieving pooled F1 scores up to 0.9381 ± 0.0005, AUROC up to 0.9976 ± 0.0001, and AUPRC up to 0.9808 ± 0.0004. Pooled statistical analyses further indicated that augmentation influenced prediction behavior beyond small numerical differences in aggregate metrics. Under comparable overall performance, BEACoN provided a more balanced false-positive and false-negative trade-off while maintaining competitive calibration behavior. Although cosine-based semantic filtering behaved similarly to standard SMOTE under the evaluated embedding setting, the findings suggest that augmentation quality in eligibility screening should be assessed not only through aggregate classification performance, but also through operational error behavior and probability reliability under severe class imbalance. Overall, the results support the potential utility of optimization-guided augmentation for screening-oriented eligibility classification. Nevertheless, the findings should be interpreted within a

Manuscript received 23 April 2026; Revised 10 June 2026; Accepted 20 June 2026; Available online 6 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1706

leakage-controlled single-corpus evaluation setting and do not yet establish generalization across independent clinical trial collections or protocol-level partitions. Additional external validation, stronger semantic filtering constraints, and alternative biomedical sentence representations remain important for assessing deployment-level robustness in future eligibility screening systems.

Acknowledgement

The authors gratefully acknowledge the support of the Center for Research Excellence and Incubation Management (CREIM), Universiti Sultan Zainal Abidin (UniSZA), Malaysia, for facilitating this research.

Funding

This research was supported by the Center for Research Excellence and Incubation Management (CREIM), Universiti Sultan Zainal Abidin (UniSZA), Malaysia.

Data Availability

The dataset used in this study was obtained from the publicly available Kaggle dataset "Clinical Trials on Cancer" provided by auriml (https://www.kaggle.com/datasets/auriml/eligibilityforcancerclinicaltrials), accessed in May 2024. The final dataset used in this study was constructed through curation and filtering of the labeledEligibilitySample1000000 corpus.

Code Availability

The code used in this study is available from the corresponding author upon reasonable request.

Author Contribution

Taslim designed the research framework, implemented the methodology, curated and analyzed the dataset, developed the models, and drafted the manuscript. Mumtazimah Mohamad supervised the study, provided methodological guidance, reviewed the evaluation protocol, and revised the manuscript.

Declarations

Ethical Approval

Not applicable. This study used publicly available clinical trial registry text and did not involve human participants, clinical interventions, or access to identifiable private patient data.

Consent for Publication

Not applicable. No individual-level participant data, images, or identifiable information are included in this manuscript.

Competing Interests

The authors declare no competing interests.

References

[1] J. Le-Rademacher, H. Gunn, X. Yao, and D. J. Schaid, "Clinical Trials Overview: From Explanatory to Pragmatic Clinical Trials," *Mayo Clin. Proc.*, vol. 98, no. 8, pp. 1241–1253, Aug. 2023, doi: 10.1016/j.mayocp.2023.04.013.

[2] S. Buccheri *et al.*, "Large simple randomized controlled trials—from drugs to medical devices: lessons from recent experience," *Trials*, vol. 26, no. 1, p. 24, 2025, doi: 10.1186/s13063-025-08724-x.

[3] X. Lu, C. Yang, L. Liang, G. Hu, Z. Zhong, and Z. Jiang, "Artificial intelligence for optimizing recruitment and retention in clinical trials: A scoping review," *J. Am. Med. Informatics Assoc.*, vol. 31, no. 11, pp. 2749–2759, 2024, doi: 10.1093/jamia/ocae243.

[4] V. Nanton *et al.*, "Boosting and broadening recruitment to UK cancer trials: towards a blueprint for action," *BMJ Oncol.*, vol. 2, no. 1, p. e000092, Nov. 2023, doi: 10.1136/bmjonc-2023-000092.

[5] K. Lee *et al.*, "Optimizing Clinical Trial Eligibility Design Using Natural Language Processing Models and Real-World Data: Algorithm Development and Validation," *JMIR AI*, vol. 3, p. e50800, 2024, doi: 10.2196/50800.

[6] O. Unlu *et al.*, "Manual vs AI-Assisted Prescreening for Trial Eligibility Using Large Language Models—A Randomized Clinical Trial," *JAMA*, vol. 333, no. 12, pp. 1084–1087, Mar. 2025, doi: 10.1001/jama.2024.28047.

[7] Q. Su, G. Cheng, and J. Huang, "A review of research on eligibility criteria for clinical trials," *Clin. Exp. Med.*, vol. 23, no. 6, pp. 1867–1879, 2023, doi: 10.1007/s10238-022-00975-1.

[8] A. Heirali *et al.*, "Eligibility Criteria of Randomized Clinical Trials in Critical Care Medicine," *JAMA Netw. Open*, vol. 8, no. 1, pp. e2454944–e2454944, Jan. 2025, doi: 10.1001/jamanetworkopen.2024.54944.

[9] C. Zihang, L. Liang, S. Qianmin, C. Gaoyi, H. Jihan, and L. Ying, "Enhanced pre-recruitment framework for clinical trial questionnaires through the integration of large language models and knowledge graphs," *Sci. Rep.*, vol. 15, no. 1, p. 27398, 2025, doi: 10.1038/s41598-025-11876-0.

[10] M. Rybinski, W. Kusa, S. Karimi, and A. Hanbury, "Learning to match patients to clinical

Manuscript received 23 April 2026; Revised 10 June 2026; Accepted 20 June 2026; Available online 6 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1706

trials using large language models," *J. Biomed. Inform.*, vol. 159, p. 104734, 2024, doi: 10.1016/j.jbi.2024.104734.

[11] S. Datta *et al.*, "AutoCriteria: A generalizable clinical trial eligibility criteria extraction system powered by large language models," *J. Am. Med. Informatics Assoc.*, vol. 31, no. 2, pp. 375–385, Feb. 2024, doi: 10.1093/jamia/ocad218.

[12] S. Gupta *et al.*, "PRISM: Patient Records Interpretation for Semantic clinical trial Matching system using large language models," *npj Digit. Med.*, vol. 7, no. 1, p. 305, 2024, doi: 10.1038/s41746-024-01274-7.

[13] K. Kantor and M. Morzy, "Machine learning and natural language processing in clinical trial eligibility criteria parsing: a scoping review," *Drug Discov. Today*, vol. 29, no. 10, p. 104139, 2024, doi: 10.1016/j.drudis.2024.104139.

[14] J. Li *et al.*, "A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. Suppl 3, p. 235, 2022, doi: 10.1186/s12911-022-01967-7.

[15] J. Park *et al.*, "Criteria2Query 3.0: Leveraging generative large language models for clinical trial eligibility query generation," *J. Biomed. Inform.*, vol. 154, p. 104649, 2024, doi: 10.1016/j.jbi.2024.104649.

[16] Y. Han, Q. Su, L. Liu, Y. Li, and J. Huang, "Structural analysis and intelligent classification of clinical trial eligibility criteria based on deep learning and medical text mining," *J. Biomed. Inform.*, vol. 160, p. 104753, 2024, doi: 10.1016/j.jbi.2024.104753.

[17] A. Bornet *et al.*, "Analysis of Eligibility Criteria Clusters Based on Large Language Models for Clinical Trial Design," *J. Am. Med. Informatics Assoc.*, vol. 32, no. 3, pp. 447–458, Mar. 2025, doi: 10.1093/jamia/ocae311.

[18] W. Kusa, O. E. Mendoza, P. Knoth, G. Pasi, and A. Hanbury, "Effective matching of patients to clinical trials using entity extraction and neural re-ranking," *J. Biomed. Inform.*, vol. 144, p. 104444, 2023, doi: 10.1016/j.jbi.2023.104444.

[19] L. Gueguen, L. Olgiati, C. Brutti-Mairesse, A. Sans, V. Le Texier, and L. Verlingue, "A prospective pragmatic evaluation of automatic trial matching tools in a molecular tumor board," *npj Precis. Oncol.*, vol. 9, no. 1, p. 28, 2025, doi: 10.1038/s41698-025-00806-y.

[20] Y. Yang, H. A. Khorshidi, and U. Aickelin, "A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems," *Front. Digit. Heal.*, vol. 6, p. 1430245, 2024, doi: 10.3389/fdgth.2024.1430245.

[21] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artif. Intell. Rev.*, vol. 57, no. 10, p. 273, 2024, doi: 10.1007/s10462-024-10884-2.

[22] A. X. Wang, V.-T. Le, H. N. Trung, and B. P. Nguyen, "Addressing imbalance in health data: Synthetic minority oversampling using deep learning," *Comput. Biol. Med.*, vol. 188, p. 109830, 2025, doi: 10.1016/j.compbiomed.2025.109830.

[23] S. F. Taskiran, B. Turkoglu, E. Kaya, and T. Asuroglu, "A comprehensive evaluation of oversampling techniques for enhancing text classification performance," *Sci. Rep.*, vol. 15, no. 1, p. 21631, 2025, doi: 10.1038/s41598-025-05791-7.

[24] O. Abdelhay, A. Shatnawi, H. Najadat, and T. Altamimi, "Resampling Methods for Class Imbalance in Clinical Prediction Models: A Scoping Review Protocol," *PLoS One*, vol. 20, no. 11, p. e0330050, 2025, doi: 10.1371/journal.pone.0330050.

[25] J. Mao, K. Huang, and J. Liu, "MLAWSMOTE: Oversampling in Imbalanced Multi-label Classification with Missing Labels by Learning Label Correlation Matrix," *Int. J. Comput. Intell. Syst.*, vol. 17, no. 1, p. 205, 2024, doi: 10.1007/s44196-024-00607-4.

[26] S. Nouas, L. Oukid, and F. Boumahdi, "Syngo: synthetic genetic oversampling technique for textual data," *Soc. Netw. Anal. Min.*, vol. 15, no. 1, p. 9, 2025, doi: 10.1007/s13278-025-01423-0.

[27] S. Ray, A. N. Sarker, N. Chatterjee, K. Bhowmik, and S. Dey, "Leveraging Large Language Models for Clinical Trial Eligibility Criteria Classification," *Digital*, vol. 5, no. 2, p. 12, 2025, doi: 10.3390/digital5020012.

[28] X. Li and Q. Liu, "A hybrid sampling algorithm for imbalanced and class-overlap data based on natural neighbors and density estimation," *Knowl. Inf. Syst.*, vol. 67, no. 3, pp. 2259–2290, 2025, doi: 10.1007/s10115-024-02281-6.

[29] V. Hernström *et al.*, "Screening performance and characteristics of breast cancer detected in the Mammography Screening with Artificial Intelligence trial (MASAI): a randomised, controlled, parallel-group, non-inferiority, single-blinded, screening accuracy study," *Lancet Digit. Heal.*, vol. 7, no. 3, pp. e175–e183, Mar. 2025, doi: 10.1016/S2589-7500(24)00267-X.

[30] auriml, "Clinical Trials on Cancer (EligibilitySample1000000.csv)." [Online].

Manuscript received 23 April 2026; Revised 10 June 2026; Accepted 20 June 2026; Available online 6 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1706

Available: https://www.kaggle.com/datasets/auriml/eligibilityforcancerclinicaltrials. Accessed: May 2024.

[31] K. raj Kanakarajan, B. Kundumani, A. Abraham, and M. Sankarasubbu, "{B}io{S}im{CSE}: {B}io{M}edical Sentence Embeddings using Contrastive learning," in *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, A. Lavelli, E. Holderness, A. Jimeno Yepes, A.-L. Minard, J. Pustejovsky, and F. Rinaldi, Eds., Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 81–86. doi: 10.18653/v1/2022.louhi-1.10.

[32] J. S. Aguilar-Ruiz and M. Michalak, "Classification performance assessment for imbalanced multiclass data," *Sci. Rep.*, vol. 14, p. 10759, 2024, doi: 10.1038/s41598-024-61365-z.

[33] A. A. Huang and S. Y. Huang, "Increasing transparency in machine learning through bootstrap simulation and shapely additive explanations," *PLoS One*, vol. 18, no. 2, p. e0281922, 2023, doi: 10.1371/journal.pone.0281922.

[34] S. Idwan, W. Etaiwi, H. Rafayia, and I. Matar, "A comprehensive review of statistical variants and enhancements of SMOTE oversampling method," *Int. J. Data Sci. Anal.*, vol. 20, pp. 6887–6904, 2025, doi: 10.1007/s41060-025-00878-w.

[35] R. Matsui, L. Guillen, S. Izumi, and T. Suganuma, "WISEST: Weighted Interpolation for Synthetic Enhancement Using SMOTE with Thresholds," *Sensors*, vol. 25, no. 24, p. 7417, 2025, doi: 10.3390/s25247417.

[36] H. Cui *et al.*, "Textual similarity calculation techniques in the medical field: a retrospective review," *Appl. Intell.*, vol. 55, no. 11, p. 814, 2025, doi: 10.1007/s10489-025-06634-8.

[37] J. Yao, X. Luo, F. Li, J. Li, J. Dou, and H. Luo, "Research on hybrid strategy Particle Swarm Optimization algorithm and its applications," *Sci. Rep.*, vol. 14, no. 1, p. 24928, 2024, doi: 10.1038/s41598-024-76010-y.

[38] R. D. Riley *et al.*, "Uncertainty of risk estimates from clinical prediction models: rationale, challenges, and approaches," *BMJ*, vol. 388, p. e080749, 2025, doi: 10.1136/bmj-2024-080749.

[39] F. M. Megahed, Y.-J. Chen, and N. Altman, "Comparing classifier performance with baselines," *Nat. Methods*, vol. 21, no. 4, pp. 546–548, 2024, doi: 10.1038/s41592-024-02234-5.

[40] K. I. Siddavatam and S. K. Shinde, "A hybrid literature review on handling imbalanced medical data: AI models and open issues," *Expert Syst. Appl.*, vol. 296, p. 129004, 2026, doi: 10.1016/j.eswa.2025.129004.

[41] M. Chen *et al.*, "Impact of human and artificial intelligence collaboration on workload reduction in medical image interpretation," *npj Digit. Med.*, vol. 7, no. 1, p. 349, 2024, doi: 10.1038/s41746-024-01328-w.

[42] Q. Jin *et al.*, "Matching patients to clinical trials with large language models," *Nat. Commun.*, vol. 15, p. 9074, 2024, doi: 10.1038/s41467-024-53081-z.

[43] C. Lin and F. Leony, "Evidence-based adaptive oversampling algorithm for imbalanced classification," *Knowl. Inf. Syst.*, vol. 66, no. 3, pp. 2209–2233, 2024, doi: 10.1007/s10115-023-01985-5.

[44] H. Chen *et al.*, "Enhancing Patient-Trial Matching With Large Language Models: A Scoping Review of Emerging Applications and Approaches," *JCO Clin. Cancer Informatics*, vol. 9, p. e2500071, 2025, doi: 10.1200/CCI-25-00071.

[45] H. Lu, L. Ehwerhemuepha, and C. Rakovski, "A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance," *BMC Med. Res. Methodol.*, vol. 22, p. 181, 2022, doi: 10.1186/s12874-022-01665-y.

[46] A. Peluso *et al.*, "Deep learning uncertainty quantification for clinical text classification," *J. Biomed. Inform.*, vol. 149, p. 104576, 2024, doi: 10.1016/j.jbi.2023.104576.

[47] Z. Premji and C. Cooper, "Same, same, but different: A method to harmonise and deduplicate study records from WHO ICTRP and ClinicalTrials.gov prior to screening," *Res. Synth. Methods*, vol. 16, no. 4, pp. 587–600, 2025, doi: 10.1017/rsm.2025.20.

Author Biography

Taslim received the B.Comp.Sc. degree in Computer Science from Universitas Putra Indonesia "YPTK", Padang, Indonesia, in 2000, and the M.Comp.Sc. degree in Computer Science from the same university in 2006. He is currently pursuing the Ph.D. degree in Computer Science at the Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin (UniSZA), Terengganu, Malaysia. He also serves as an academic staff member at Universitas Lancang Kuning, Pekanbaru, Indonesia. His research interests include artificial intelligence and data analytics, particularly in healthcare applications. In 2021, he received a DIKTI-

Manuscript received 23 April 2026; Revised 10 June 2026; Accepted 20 June 2026; Available online 6 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1706

funded research grant (Hibah Riset Kemanusiaan PTS) focusing on predictive modeling for COVID-19 patient outcomes, including length-of-stay estimation. He is an active member of professional bodies, including the International Association of Engineers (IAENG).

Mumtazimah Mohamad received the Bachelor's degree in Information Technology from Universiti Kebangsaan Malaysia in 2000 and the M.Sc. degree in Computer Science from Universiti Putra Malaysia in 2004. She received the Ph.D. degree in Computer Science from Universiti Malaysia Terengganu in 2014. She began her academic career as a Junior Lecturer in 2000 and is currently an Associate Professor with the Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin (UniSZA), Terengganu, Malaysia. Her academic roles include teaching, research, and academic governance. Her research interests include pattern recognition, artificial intelligence, machine learning, data mining, and parallel processing. She has authored numerous publications and actively serves as a reviewer and technical committee member for reputable journals and conferences.

Manuscript received 23 April 2026; Revised 10 June 2026; Accepted 20 June 2026; Available online 6 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1706