IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 4, August 2026, pp. 3944∼3954
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i4.pp3944-3954

Intravenous immunoglobulin resistance prediction in Kawasaki disease using oversampled transformer embeddings

Namitha Thattarassery Nanappan[1,2], Raghavendra Srinivasaiah[3], Vinith Rejathalal[4]
[1]Department of Computer Science and Engineering, CHRIST (Deemed to be University), Bengaluru, India
[2]Department of Computer Science and Engineering, Mahaguru Institute of Technology, Kayamkulam, India
[3]Department of AI and Data Science Engineering, CHRIST (Deemed to be University), Bengaluru, India
[4]Department of Artificial Intelligence, Amrita Vishwa Vidyapeetham, Coimbatore, India

Article Info

Article history:

Received Jun 20, 2025
Revised Jul 11, 2026
Accepted Jul 21, 2026

Keywords:

Intravenous immunoglobulin resistance
Kawasaki disease
Machine learning
Oversampling
Transformer embeddings

ABSTRACT

Kawasaki disease (KD) is a leading cause of acquired heart disease in children under five. Although intravenous immunoglobulin (IVIG) treatment is usually effective, 10–20% of cases are resistant and at higher risk for coronary complications. Early prediction of IVIG resistance is critical but difficult due to the rarity of KD and imbalanced clinical data. To address this, we propose a novel technique called sentence transformer embeddings with synthetic minority over-sampling technique (SMOTE) oversampling (STESO), which leverages the complementary strengths of transformer-based representation learning and synthetic oversampling. Pretrained models such as paraphrase-MiniLM-L3-v2 are used to convert tabular clinical data into dense text-based embeddings, capturing deeper semantic relationships across features. By coupling these rich embeddings with SMOTE, we balance class distributions directly in the semantic space, enabling traditional machine learning (ML) models to more effectively detect minority (resistant) cases. This synergy yielded substantial improvements in sensitivity and F1-score, with random forest (RF) combined with STESO (RF-STESO) achieving the highest overall performance. Among the models evaluated, our proposed model attained best result as accuracy of 0.85, sensitivity of 0.81, specificity of 0.89, and F1-score of 0.85. Our results underscore that the joint use of transformer embeddings and oversampling is more effective than either approach in isolation, offering a promising pathway for rare disease prediction tasks such as IVIG resistance prediction in KD.

Corresponding Author:

Namitha Thattarassery Nanappan
Department of Computer Science and Engineering, CHRIST (Deemed to be University)
Bengaluru, Karnataka, India
Email:namitha.tn@res.christuniversity.in

1. INTRODUCTION

Kawasaki disease (KD), first described by Tomisaku Kawasaki in 1967 and published in 1974, is a leading cause of acquired heart disease in children worldwide. Children under the age of five are the most affected. They experience prolonged fever, conjunctival injection, swollen mucosa, swollen lymph nodes, a characteristic rash, and mucosal Inflammation. The cause of KD continues to be unanswerable, and the absence of treatment for KD can and will lead to severe long-term consequences. Coronary artery aneurysm (CAA),

myocarditis, and arrhythmias, to name a few, are all complications that can be developed and/or exacerbated if KD is not treated [1]. Japan has the highest incident rate of KD, with about 12,000 cases each year. The primary KD treatment has been and continues to be intravenous immunoglobulin (IVIG), and the coronary complications will be reduced with this treatment. Upon treatment, some patients demonstrate persistent or recurring fever 36 hours after the infusion. These patients at this point demonstrate IVIG treatment resistance. Risk of coronary complications for these patients is around 9 times more likely than for those that respond to treatment [2]. The main treatment of KD is IVIG, and the most complicated portion of the treatment is predicting treatment resistance. The unbalanced datasets for KD have overwhelmingly more responsive cases than cases resistant to treatment.

Early identification of IVIG resistance is critical to provide additional interventions, such as corticosteroids or infliximab, which can significantly improve patient outcomes [3]. Predicting IVIG resistance remains challenging due to the imbalanced nature of KD datasets, where IVIG-responsive cases dominate. Sensitivity scores of existing systems such as Kobayashi, Egami, and Sano scores fall under 50%, compromising these systems overall predictive capability [4]. Ongoing research efforts in this area are summarized in [5] and developed a knowledge framework regarding IVIG resistance predictions. Mirata *et al*. [6] put together a review which suggested that a large portion of present-day research focused on classical machine learning (ML) methods.

There are classical ML techniques which include logistic regression (LR), random forest (RF), support vector machines (SVM), gradient boosting (GB), light gradient boosting machine (LightGBM), extreme gradient boosting (XGBoost), and neural networks (NN). For example, Lam *et al.* [1] performed an evaluation of the models which included LR, RF, naive Bayes, gradient boosting machine (GBM), and NN and documented fairly low sensitivity with LR scoring 0.216, and SVM 0.60. Sunaga *et al*. [7] showed implementation of LightGBM and documented sensitivity of 0.60. The study, [8] incorporated LR, while [9] utilized a wide variety of models that consisted of LR, multi-layer perceptron (MLP), RF, categorical boosting (CatBoost), explainable boosting machine (EBM), and GBM.

The papers [10] and [11] both directed their efforts on models that use RF. In [12] XGBoost model showed strong performance on the testing set where they scored an area under the curve (AUC) of 0.821, an accuracy of 74.8%, sensitivity of 88.9%, and specificity of 68.3%. Also, Liu *et al.* [13] used LR, SVM, XGBoost, and LightGBM. Even with the range of models and the growing volume of datasets, most of the approaches continue to have poor sensitivity which is a serious problem given the class imbalance present in KD datasets. Although performance metrics such as accuracy, specificity, and AUC are commonly reported, enhancing sensitivity remains essential for making these prediction models clinically meaningful.

This study integrates ML with transfer learning and innovative natural language processing technologies to find a solution. In clinical natural language processing, transformers are becoming essential technologies and are particularly and highly versatile at understanding dense and intricate medical texts. Most recent research uses domain-specific models that are trained on healthcare materials and surpass general-purpose models for diagnosis coding and symptom extraction. For instance, NorDeClin-bidirectional encoder representations from transformers (BERT) trained on Norwegian clinical notes and significantly enhanced results on international classification of diseases (ICD) classification [14].

Spanish transformers, for instance, enhanced extraction of heart failure symptoms from electronic health records (EHRs) [15]. It is no longer an academic exercise to fine-tune generative pre-trained transformers (GPTs) on medical corpora for clinical summarization and diagnosis support, sharpening the focus on the clinical specialization of healthcare natural language processing [16]. Unlike the traditional hand-coded models, wherein numerical data are represented as texts, language models, and more particularly sentence transformers, are used to find the complex texts that describe the relationship between the data. Class imbalance is corrected using the synthetic minority over-sampling technique (SMOTE), and the enhanced embeddings are passed to the classical ML methods.

This approach bridges the gap between conventional scoring systems and modern artificial intelligence (AI) models, resulting in significant improvements in prediction sensitivity and overall performance. This work introduces sentence transformer embeddings with SMOTE oversampling (STESO) algorithm as a novel framework for early and accurate identification of IVIG-resistant KD patients. By situating STESO within the broader paradigm of clinical decision support systems (CDSS), this study shows how sentence-transformer embeddings can repurpose structured clinical data for sensitive prediction tasks, while laying the groundwork for future integration with narrative records.

2. DATA AND METHODS

The dataset used in this study, as described in [13], consists of 1,398 medical records of KD patients covering the years 2015-2020. The records include 1,240 cases of IVIG-responsive patients and 158 cases of IVIG-resistant patients, and contained 31 variables of demographic, clinical, and imaging, and laboratory data. To tackle the issue of data imbalance and to concentrate on impactful predictors, as described in the original study, the dataset underwent feature selection to determine relevant predictors using univariate analysis via the chi-square test in SPSS version 25.0 (p-value 0.05) [13].

For the purposes of ensuring consistency and generalizability of the approach, we used the dataset in the research [17]. Since no predefined feature subset was available for this dataset, we employed Shapley additive explanations (SHAP) analysis on the full set of variables to identify the top 14 features contributing most to IVIG resistance predictions. SHAP analysis highlighted features such as day of fever and platelet count (PLT), both established clinical markers in KD, thereby supporting the biological plausibility of our approach. To make the dataset compatible with a language transformer model, numerical features were transformed into ordinal categorical values by analyzing central tendency, spread, and quartile metrics. For instance, continuous variables were categorized into levels, ensuring consistency across all selected features.

The processed categorical data was further converted into structured sentences suitable for the sentence transformer [18]. One record yielded the following clinical summary. The patient had high days of illness. Lymphocyte count was lymphocyte medium, and hemoglobin (HB) levels were HB medium. Neutrophil percentage was neutr very very high, with an neutrophil-to-lymphocyte ratio (NLR) of neutr ultimate maximum. Globulin was globulin high while albumin was albumin minimum. Aspartate aminotransferase (AST) levels were AST less, and serum sodium was serum sodium very very high. Total bilirubin (TBIL) was TBIL medium, procalcitonin (PCT) was PCT maximum, and alanine aminotransferase (ALT) was ALT medium.

PLT levels were PLT less, and gamma-glutamyl transferase (GGT) was GGT minimum. This transformation allowed the model to handle the dataset effectively while maintaining the integrity of critical information. These category thresholds were cross-checked against both statistical distribution patterns and established clinical relevance (e.g., low PLTs and prolonged fever are well-documented risk factors for IVIG resistance), thereby ensuring that the semantic transformation preserved medical plausibility. These preprocessing steps ensured the dataset was optimized for the transformer model and highlighted the importance of robust preprocessing in predictive modeling for rare diseases like KD. As shown in Figure 1, advanced ML techniques are still underutilized in this field. Our proposed method bridges this gap by using transformers as the base model.

The method section describes in detail the underlying models-BERT, sentence-BERT (SBERT), and sentence transformer fine-tuning (SetFit)—before turning to the introduction of our new model, STESO. This ordering allows for a well-founded comprehension of the progression of the proposed architecture. BERT is a transformer-based language model designed for various tasks, such as text classification. It uses a bidirectional encoder and is pre-trained using self-supervised learning tasks, including masked language modeling (MLM) and next sentence prediction (NSP) [19]. These tasks allow BERT to capture the structural and semantic intricacies of human language without requiring labeled data. The MLM task randomly masks 15% of tokens in the input and predicts them based on the surrounding context. The NSP task trains the model to determine whether a given sentence follows another logically. Architecturally, BERT consists of stacked transformer encoders. The base model (BERTBASE) has 12 layers, 768 hidden dimensions, and 12 self-attention heads, with 110M parameters in total, while the large model (BERTLARGE) has 24 layers and 340M parameters. Despite its effectiveness in numerous natural language processing tasks, BERT embeddings underperform in pairwise sentence tasks, such as clustering and similarity detection. To address this limitation, SBERT was proposed.

Sentence transformer (SBERT) extends BERT by using a Siamese or triplet network architecture to handle pairwise or triplet-based tasks such as classification and similarity comparison [20]. The model fine-tunes BERT embeddings to generate sentence-level representations that are more effective for pairwise and clustering tasks [21]. In its architecture, pairs or triplets of sentences are passed through the network simultaneously, generating embeddings that can be directly used for tasks like semantic similarity or classification. In cases where there is a need for quick comparisons, SBERT embeddings come in handy, especially for downstream tasks featuring sentence representations. SetFit expands on the concept of SBERT by applying a two-phase training methodology for few shot learning [22]. For some recent work, few shot

learning has been targeted in the medical field which has data constrained medical tasks specifically in rare diseases. For instance, 3D few-shot learning was developed for a model that classifies rare knee injuries in magnetic resonance imaging (MRI) [23] which showcases generalization to low-resource imaging tasks, a challenge in the field of medical imaging.

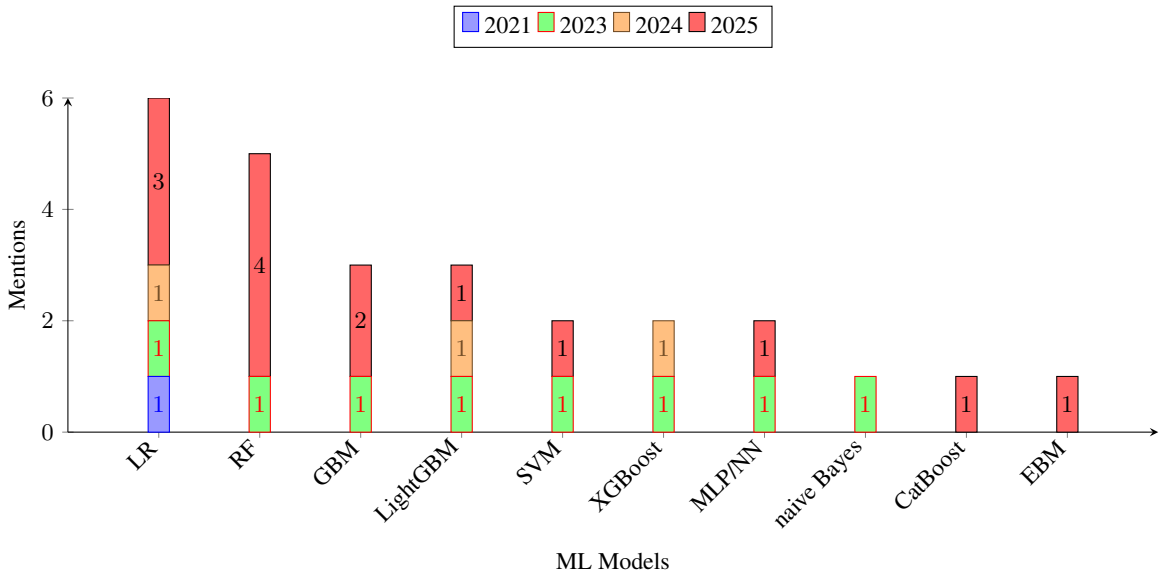

Figure 1. Model mentions across years

Furthermore, generative data augmentation has been shown to improve performance in diagnostic visual recognition of few-shot cases which improves the recognition of certain conditions that are more difficult to diagnose [24]. In a similar design, large language models were used with ontologies in a rare pediatric epilepsy case [25] which provided a few-shot extraction of information regarding drug efficacy which shows the enhanced clinical natural language processing task of semantic integration. In the first phase, the training of the model uses a contrastive technique where sentence pairs are made into positive and negative examples. This process significantly expands the training dataset size, scaling it quadratically to $O(K^2)$, where K is the original sample size. This phase fine-tunes the sentence transformer to effectively distinguish between contrasting sentences. In the second phase, embeddings generated by the fine-tuned transformer are used to train a downstream classifier with labeled data. This combination of contrastive learning and classification makes SetFit particularly effective in data-scarce environments, offering a robust and efficient foundation for generating embeddings tailored to specific tasks. Figure 2 illustrates the overall training and inference process of the SetFit model.

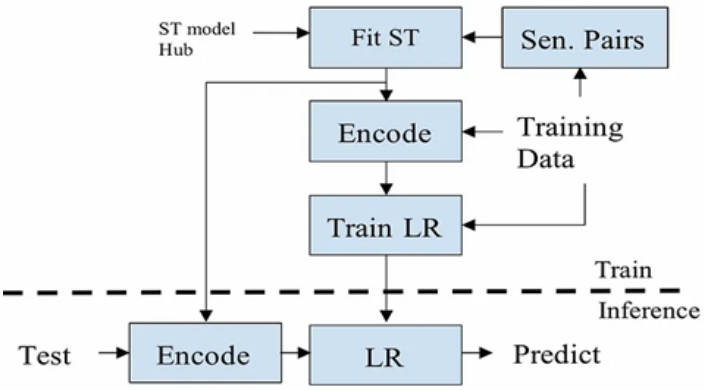

Figure 2. SetFit training and inference block diagram [22]

2.1. Proposed sentence transformer embeddings with SMOTE oversampling methodology

Even though BERT and other pre-trained language transformer models create sentence embeddings, they do not detect subtle differences especially in imbalanced datasets. Here, the gap is addressed by using SMOTE in the embedding space for the minority embeddings helps classifiers work better. The STESO methodology combines sentence transformer embeddings, SMOTE oversampling, and ordinary ML classifiers to manage the extreme imbalance in class distribution. To align with transformer models, we converted numerical features to textual sentences. These sentences are then encoded and run through the paraphrase-MiniLM-L3-v2 model from sentence transformers, and the embeddings are created by averaging the last hidden layer across all tokens.

This creates high-dimensional embeddings, or vector representations, e_i, that capture various semantic and contextual aspects. These embeddings go beyond simple encoding by higher-order interaction modeling with SHAP-selected predictors, like characterizing prolonged fever and low PLT as a single clinical indicator. This allows the model to approximate clinician-style reasoning, improves generalization on imbalanced datasets, and maintains traceability to established predictors while enriching them with contextual meaning. Let the dataset be represented by (1).

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \tag{1}$$

In (1), where each $x_i \in \mathbb{R}^{14}$ is a 14-dimensional numeric feature vector, and $y_i \in \{0, 1\}$ is the associated binary class label.

Each feature vector $x_i = [x_{i1}, x_{i2}, \ldots, x_{i14}]$ is transformed into a natural language sentence. This transformation is expressed through (2).

$$S_i = \text{concat}(\text{feature1 is } x_{i1}, \text{feature2 is } x_{i2}, \ldots, \text{feature14 is } x_{i14}) \tag{2}$$

As defined in (2), allowing textual interpretation of numerical features. This sentence is passed through the pretrained sentence transformer model (STM) T_{Sent}. The model tokenizes the input and encodes it into contextualized token-level embeddings are given by (3).

$$\text{Encoder}(S_i) = [h_1, h_2, \ldots, h_m] \tag{3}$$

Where $h_j \in \mathbb{R}^d$ and m is the number of tokens. These token embeddings are aggregated using mean pooling to obtain a fixed-size vector as in (4).

$$e_i = \frac{1}{m} \sum_{j=1}^{m} h_j, \quad \text{where } e_i \in \mathbb{R}^d \tag{4}$$

All embeddings are then stacked to form the matrix. Stacking all embeddings forms the matrix in (5).

$$E = \begin{bmatrix} e_1^T \\ e_2^T \\ \vdots \\ e_n^T \end{bmatrix} \in \mathbb{R}^{n \times d} \tag{5}$$

To handle class imbalance, SMOTE is applied in the embedding space. The minority class embeddings are defined as (6).

$$E^+ = \{e_i \in \mathbb{R}^d \mid y_i = 1\} \tag{6}$$

Each $e_i \in E^+$ is used to find k nearest neighbors using Euclidean distance. The nearest neighbors are identified as in (7).

$$\text{NN}(e_i) = \{e_{i1}, e_{i2}, \ldots, e_{ik}\} \subset E^+ \quad \text{such that } \|e_i - e_{ij}\|_2 \text{ is minimized} \tag{7}$$

New synthetic samples are generated through interpolation as shown in (8).

$$e_{\text{new}} = e_i + \lambda \cdot (e_j - e_i), \quad \lambda \sim \mathcal{U}(0, 1) \tag{8}$$

The resulting balanced dataset and corresponding labels are provided in (9) and (10).

$$E_{\text{SMOTE}} = E \cup \{e^1_{\text{new}}, e^2_{\text{new}}, \ldots, e^m_{\text{new}}\} \tag{9}$$

$$y_{\text{SMOTE}} = y \cup \{\underbrace{1, 1, \ldots, 1}_{m}\} \tag{10}$$

The embeddings E_{SMOTE} and labels y_{SMOTE} can now be used to train classical ML models. If LR is chosen, the predicted probabilities $\hat{y}_i \in [0, 1]$ are optimized using the binary cross-entropy loss. The STESO architecture is detailed in Algorithm 1. This study takes advantage of the ability of pretrained sentence transformers to generate meaningful sentence embeddings from minimal data. These embeddings are then balanced using SMOTE to address class imbalance. By combining this with traditional classifiers, STESO offers a practical solution for achieving strong performance even when data is limited.

Algorithm 1 STESO: sentence transformer embeddings with SMOTE oversampling

Require: tabular dataset $D = \{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^{14}$, $y_i \in \{0, 1\}$; pretrained sentence transformer T_{sent}; SMOTE function.
Ensure: trained classifier $\mathcal{L}$ and evaluation metrics.
1: **Sentence Conversion:** for each x_i, convert to clinical sentence S_i
2: **Sentence Embedding:** $e_i = T_{\text{sent}}(S_i)$
3: Form embedding matrix: $E = [e_1, ..., e_n]^T \in \mathbb{R}^{n \times d}$
4: **Oversampling train set:** $(E', y') = \text{SMOTE}(E, y)$
5: **Training:** train a classifier $\mathcal{L}$ (e.g., RF, LR) on (E', y')
6: **Evaluation:** compute accuracy, sensitivity, specificity, and F1-score on test set

3. RESULTS AND DISCUSSION

This section compares traditional ML, deep learning (DL), SetFit-based few-shot learning with a STM, and the proposed STESO framework. Results show progressive improvements across methods. On the unbalanced dataset of 1,398 patient records, traditional ML models (LR, RF, SVM, k-nearest neighbor (KNN)) and a feed-forward neural network (FFNN) achieved high accuracy and specificity but performed poorly in predicting IVIG resistance, with low sensitivity and F1-scores as in Table 1.

Table 1. Performance metrics of conventional ML and DL models using dataset [13]

Metric	LR	RF	SVM	KNN	FFNN	STM
Accuracy	0.92	0.93	0.90	0.90	0.86	0.88
Sensitivity	0.36	0.48	0.15	0.24	0.42	0.48
Specificity	0.99	0.93	1.00	0.93	0..92	0.93
F1-score	0.53	0.70	0.41	0.37	0.46	0.47

To evaluate sentence transformer, this study fine-tuned the SetFit framework using its pretrained sentence transformer backbone on balanced datasets with increasing shot sizes (25, 50, 100, 158). Sensitivity improved steadily, reaching 0.709 with 158-shot samples (Figure 3). Since large datasets are uncommon in rare diseases such as KD, this study extended this idea by directly employing the same STM to generate semantically rich embeddings from the full clinical dataset. Applying SMOTE in this embedding space produced balanced training data that preserved clinical structure, enabling traditional ML models to more effectively detect resistant cases. As shown in Figures 4 and 5, STESO consistently improved sensitivity and F1-scores, with RF achieving the best overall performance due to its capacity to capture complex non-linear interactions within the enriched feature space. These findings underscore STESO's strength as a scalable solution for predictive modeling in rare and imbalanced medical datasets.

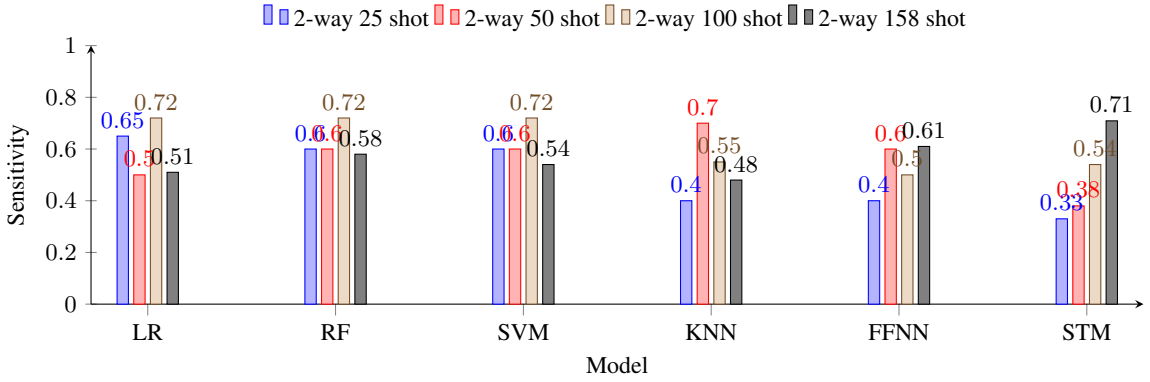

Figure 3. Sensitivity improvement of STM with increasing shot counts

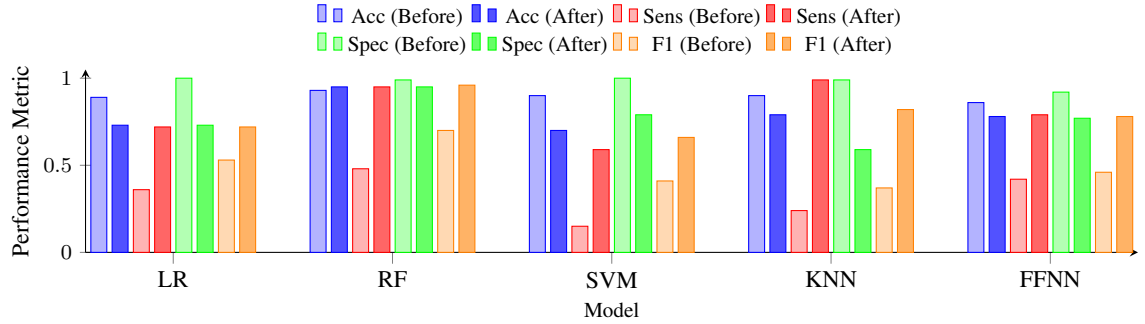

Figure 4. Performance comparison of selected ML models before and after applying sentence transformer embeddings with SMOTE. RF-STESO shows the best overall performance across all metrics

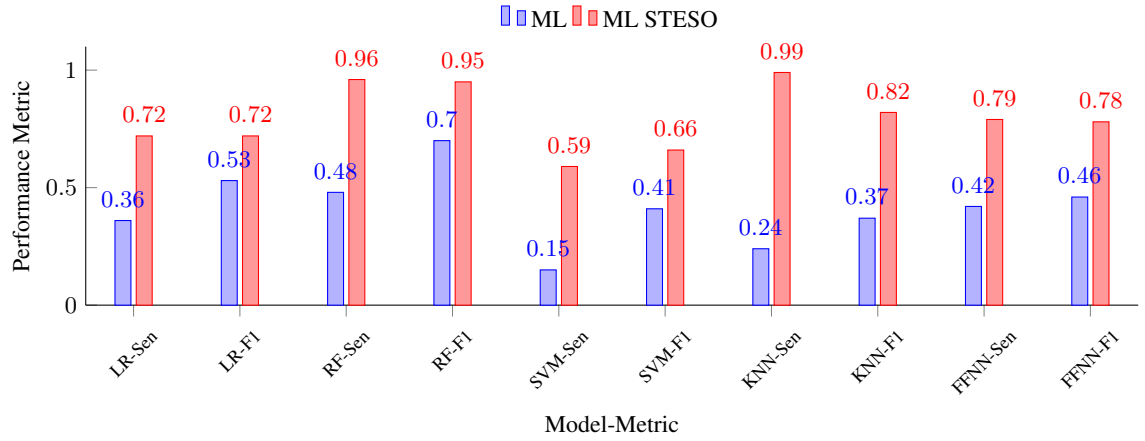

Figure 5. Improvement in sensitivity and F1-score with ML STESO

To evaluate robustness beyond the initial dataset, this study applied STESO to a second highly imbalanced dataset [17]. As shown in Table 2, STESO improved performance across all models (LR, RF, SVM, KNN, and FFNN). These gains stem from the synergy of transformer-based embeddings, which capture meaningful clinical patterns, and SMOTE, which balances minority cases in the semantic space. This combination enabled better detection of resistant cases compared to baselines. All improvements were confirmed as statistically significant using McNemar's test ($p < 0.05$) with 10 fold cross validation with RF and RF-STESO, demonstrating STESO's robustness and adaptability for imbalanced datasets.

Table 2. Sensitivity and F1-score improvements on dataset 2 [17] (ML vs ML STESO) to ensure model consistency and generalizability

Model	Sensitivity (Before)	Sensitivity (After)	F1-score (Before)	F1-score (After)
LR	0.14	0.52	0.24	0.32
RF	0.14	0.81	0.24	0.85
SVM	0.00	0.56	0.00	0.35
KNN	0.10	0.60	0.17	0.31
FFNN	0.35	0.46	0.26	0.60

According to the Table 3, the comparison shows that RF-STESO is superior to both convolutional neural networks (CNNs) which use convolutional filters to capture local feature interactions, and the feature tokenizer transformer (FT-Transformer) which is an attention-based model for tabular data, providing greater and more balanced values across all the metrics of accuracy, sensitivity, specificity, and F1-score. Even though CNNs deliver acceptable performance and FT-Transformers yield disproportionate outcomes across the metrics, RF-STESO still shows remarkable consistency across both datasets. This result underscores the power of integrating semantic embeddings with RF to achieve dependable forecasting for RF-STESO in clinically imbalanced scenarios.

Table 3. Performance comparison of CNN, FT-Transformer, and RF STESO across datasets

Model/Dataset	Accuracy	Sensitivity	Specificity	F1-score
CNN [13]	0.89	0.83	0.83	0.83
FT-Transformer[13]	0.90	0.27	1.00	0.43
RF STESO [13]	0.95	0.96	0.95	0.96
CNN [17]	0.83	0.93	0.74	0.82
FT-Transformer [17]	0.80	0.32	0.94	0.41
RF-STESO [17]	0.85	0.81	0.89	0.85

3.1. Ablation study, failure analysis, and future directions

To evaluate each component of embedding generation and oversampling, in this ablation study tested five scenarios: i) ML models trained on raw clinical data, ii) ML models trained on raw clinical data after SMOTE oversampling, iii) ML models trained on sentence-transformer embeddings with SMOTE oversampling, which is our proposed STESO framework, iv) a comparative benchmark of STESO against DL approaches, CNN and FT-Transformer, which was trained directly on tabular data as shown in Tables 3 and 4, and v) oversampled FT Transformer embedding with sentence transformer embeddings as shown Table 5. Performance evaluation showed that STESO resulted in the highest sensitivity and F1-score. This illustrates the synergistic effect that learning semantic representations and adjusting class distributions on detecting the minority class in IVIG resistance prediction.

This model is a novel decision-support conceptual tool that helps clinicians in the early identification of IVIG resistance. Such a model would enable clinicians to identify high-risk patients that require close monitoring, timely echocardiographic evaluation, or the consideration of adjunct therapies. However it's use is constrained by limitations such as the relatively small dataset, site-specific bias, and the need for external validation before clinical deployment.

In proposed method, some models showed a decline in accuracy and specificity. But in our case, the accurate prediction of IVIG resistance is more critical than predicting responsive patients. So, we mainly focused on improving the sensitivity and F1-score. Other oversampling approaches like adaptive synthetic (ADASYN) also not explored in current study. Eventhough, the embeddings enable the representation learning, these latent features are difficult to interpret. To improve interpretation, explainable techniqus like SHAP can be utilized in future.

Future work will also include external validation on geographically distinct cohorts to enhance generalizability, alongside efforts to improve interpretability of transformer embeddings, as well as exploring recent advances in multimodal fusion and graph-based modeling to capture richer clinical representations. While embeddings bridge structured and unstructured clinical data for CDSS, ethical and deployment considerations remain critical. This study plan to address bias mitigation, strengthen data privacy safeguards, conduct multi-center clinical validation, and explore integration into EHR systems and mobile diagnostics to support real-world adoption.

Table 4. Comparison of LR, RF, SVM, and KNN under different data balancing strategies using data [17]

Method	LR	RF	SVM	KNN
Sensitivity				
ML models (no SMOTE)	0.14	0.14	0.00	0.10
ML + SMOTE (raw features)	0.67	0.25	0.39	0.50
Sentence transformer embeddings + SMOTE + ML	0.52	0.81	0.56	0.60
F1-score				
ML models (no SMOTE)	0.24	0.24	0.00	0.17
ML + SMOTE (raw features)	0.44	0.29	0.34	0.33
Sentence transformer embeddings + SMOTE + ML	0.32	0.85	0.35	0.31

Table 5. Comparison of sentence transformer and FT-Transformer embeddings with SMOTE for selected ML models using data [17]

Model	Sensitivity (Sentence transformer)	Sensitivity (FT-Transformer)	F1-score (Sentence transformer)	F1-score (FT-Transformer)
LR	0.52	0.25	0.32	0.30
RF	0.81	0.25	0.85	0.36
SVM	0.56	0.37	0.35	0.42
KNN	0.60	0.37	0.31	0.37

4. CONCLUSION

STESO demonstrates that combining sentence transformer-based embeddings with oversampling in the semantic space provides a powerful solution for IVIG resistance prediction in KD. By leveraging pretrained sentence transformers, the proposed method in this study captures deeper clinical relationships, and applying oversampling in this enriched representation space generates synthetic resistant cases that preserve semantic and biological plausibility, unlike raw feature oversampling which can produce unrealistic combinations. This synergy enhances sensitivity and F1-score across models, with RF–STESO achieving the strongest performance by exploiting complex interactions within the embeddings. Moreover, the embedding-space oversampling introduces controlled perturbations akin to regularization, improving robustness and generalization to unseen data. Beyond KD, this strategy offers a transferable pathway for other rare disease prediction tasks where limited and imbalanced datasets remain a critical challenge, making transformer–oversampling integration a promising step toward more reliable and personalized healthcare solutions.

FUNDING INFORMATION

Authors state there is no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Namitha Thattarassey Nanappan	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Raghavendra Srinivasaiah		✓				✓		✓	✓	✓	✓	✓		
Vinith Rejathalal	✓		✓	✓		✓			✓		✓		✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal Analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject Administration
Fu : **Fu**nding Acquisition

CONFLICT OF INTEREST STATEMENT

The author declares that there are no known conflicts of interest associated with this publication. There are no financial or personal relationships that could inappropriately influence or bias the content of this work.

INFORMED CONSENT

Not applicable. This study did not involve human participants, human data, or any personally identifiable information. All data used were either publicly available, fully anonymized, or derived from non-human sources, and therefore no informed consent was required from individuals.

ETHICAL APPROVAL

Not applicable. This research did not involve human subjects, human biological materials, or experimental procedures on animals. The work was conducted solely on computational models, publicly available datasets, or non-sensitive data that did not require intervention with living organisms. Therefore, ethical approval from an institutional review board or animal ethics committee was not necessary for this study.

DATA AVAILABILITY

The dataset used in this study is publicly available and was obtained from previously published research as cited in references [13] and [17]. No new data were generated for this work.

REFERENCES

[1] J. Y. Lam *et al*., "Intravenous immunoglobulin resistance in Kawasaki disease patients: prediction using clinical data," *Pediatric Research*, vol. 95, no. 3, pp. 692–697, Feb. 2024, doi: 10.1038/s41390-023-02519-z.

[2] Y. Wang *et al*., "Development of an immunoinflammatory indicator-related dynamic nomogram based on machine learning for the prediction of intravenous immunoglobulin-resistant Kawasaki disease patients," *International Immunopharmacology*, vol. 134, Jun. 2024, doi: 10.1016/j.intimp.2024.112194.

[3] X. Wang, X. Shi, X. Guo, S. Chen, X. Lin, and F. Yang, "Effectiveness of initial corticosteroid treatment in Kawasaki disease children suspected to be IVIG resistant," *Pediatric Cardiology*, vol. 46, no. 8, pp. 2315–2321, Dec. 2025, doi: 10.1007/s00246-024-03657-9.

[4] U. K. Akca *et al*., "Comparison of IVIG resistance predictive models in Kawasaki disease," *Pediatric Research*, vol. 91, no. 3, pp. 621–626, Feb. 2022, doi: 10.1038/s41390-021-01459-w.

[5] J. Zhang *et al*., "Knowledge framework of intravenous immunoglobulin resistance in the field of Kawasaki disease: a bibliometric analysis (1997-2023)," *Immunity, Inflammation and Disease*, vol. 12, no. 5, May 2024, doi: 10.1002/iid3.1277.

[6] D. Mirata *et al*., "Learning-based models for predicting IVIG resistance and coronary artery lesions in Kawasaki disease: a review of technical aspects and study features," *Pediatric Drugs*, vol. 27, no. 4, pp. 465–479, Jul. 2025, doi: 10.1007/s40272-025-00693-7.

[7] Y. Sunaga *et al*., "A simple scoring model based on machine learning predicts intravenous immunoglobulin resistance in Kawasaki disease," *Clinical Rheumatology*, vol. 42, no. 5, pp. 1351–1361, May 2023, doi: 10.1007/s10067-023-06502-1.

[8] S. Wang *et al*., "Establishment and validation of risk prediction model to predict intravenous immunoglobulin-resistance in Kawasaki disease based on meta-analysis of 15 cohorts," *Italian Journal of Pediatrics*, vol. 51, no. 1, Feb. 2025, doi: 10.1186/s13052-025-01889-w.

[9] E. J. Cheon, G. B. Kim, and S. Park, "Predictive modeling of consecutive intravenous immunoglobulin treatment resistance in Kawasaki disease: a nationwide study," *Scientific Reports*, vol. 15, no. 1, Jan. 2025, doi: 10.1038/s41598-025-85394-4.

[10] Y. Xia *et al*., "A machine learning-based model to predict intravenous immunoglobulin resistance in Kawasaki disease," *iScience*, vol. 28, no. 3, Mar. 2025, doi: 10.1016/j.isci.2025.112004.

[11] Y. He *et al*., "Interpretable web-based machine learning model for predicting intravenous immunoglobulin resistance in Kawasaki disease," *Italian Journal of Pediatrics*, vol. 51, no. 1, Jun. 2025, doi: 10.1186/s13052-025-02036-1.

[12] L. Deng *et al*., "Construction and validation of predictive models for intravenous immunoglobulin–resistant Kawasaki disease using an interpretable machine learning approach," *Clinical and Experimental Pediatrics*, Jul. 2024, doi: 10.3345/cep.2024.00549.

[13] J. Liu *et al*., "A machine learning model to predict intravenous immunoglobulin-resistant Kawasaki disease patients: a retrospective study based on the chongqing population," *Frontiers in Pediatrics*, vol. 9, Nov. 2021, doi: 10.3389/fped.2021.756095.

[14] P. D. Ngo *et al*., "Domain-specific pretraining of NorDeClin-bidirectional encoder representations from transformers for international statistical classification of diseases, tenth revision, code prediction in Norwegian clinical texts: model development and evaluation study," *JMIR AI*, vol. 4, pp. e66153–e66153, Aug. 2025, doi: 10.2196/66153.

[15] J. Mata, V. Pachón, A. Manovel, M. J. Maña, and M. D. L. Villa, "Multicriteria optimization of language models for heart failure with preserved ejection fraction symptom detection in Spanish electronic health records: comparative modeling study," *Journal of Medical Internet Research*, vol. 27, Jul. 2025, doi: 10.2196/76433.

[16] Y. H. Lim, P. S. Q. Yeoh, and K. W. Lai, "Evaluating fine-tuned GPT models on different datasets in the healthcare domain," *Innovation and Emerging Technologies*, vol. 12, Jan. 2025, doi: 10.1142/S2737599425500124.

[17] T. Wang, G. Liu, and H. Lin, "A machine learning approach to predict intravenous immunoglobulin resistance in Kawasaki disease patients: a study based on a Southeast China population," *PLoS ONE*, vol. 15, no. 8, Aug. 2020, doi: 10.1371/journal.pone.0237321.

[18] L. Tunstall *et al*., "Efficient few-shot learning without prompts," 2022, *arXiv: 2209.11055*.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.
[20] N. Reimers and I. Gurevych, "Sentence-BERT: sentence embeddings using siamese BERT-networks," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019, pp. 3982-3992.
[21] L. Stankevičius and M. Lukoševičius, "Extracting sentence embeddings from pretrained transformer models," *Applied Sciences*, vol. 14, no. 19, Oct. 2024, doi: 10.3390/app14198887.
[22] M. Wasserblat, "Sentence transformer fine-tuning (SetFit)," community.intel.com. Accessed: Jul. 30, 2026. [Online]. Available: https://community.intel.com/t5/Blogs/Tech-Innovation/Artificial-Intelligence-AI/Sentence-Transformer-Fine-Tuning-SetFit/post/1407712
[23] V. H. Dang *et al*., "Exploration of 3D few-shot learning techniques for classification of knee joint injuries on MR images," *Diagnostics*, vol. 15, no. 14, Jul. 2025, doi: 10.3390/diagnostics15141808.
[24] X. Wang, Z. Chu, and L. Zhu, "Research on data augmentation algorithms for few-shot image classification based on generative adversarial networks," *Academia Nexus Journal*, vol. 3, no. 3, Oct. 2024.
[25] P. Golnari *et al*., "Ontology accelerates few-shot learning capability of large language model: a study in extraction of drug efficacy in a rare pediatric epilepsy," *International Journal of Medical Informatics*, vol. 201, Sep. 2025, doi: 10.1016/j.ijmedinf.2025.105942.

BIOGRAPHIES OF AUTHORS

Namitha Thattarassery Nanappan is a research scholar at the Department of Computer Science and Engineering, CHRIST (Deemed to be University), India. She completed her master's degree in Computer Science and Engineering from Anna University, Chennai, India, in 2013. Her areas of interest include ML, federated learning, and rare disease research. She is currently pursuing her research in these domains, focusing on developing innovative solutions for rare disease prediction and management using advanced ML techniques. She has presented her work at various national and international conferences and has published her research in reputed journals. She can be contacted at email: namitha.tn@res.christuniversity.in.

Raghavendra Srinivasaiah is currently working as associate professor in the Department of AI and Data Science Engineering at CHRIST (Deemed to be University), Bengaluru, India. He completed his Ph.D. degree in Computer Science and Engineering from Visvesvaraya Technological University, Belgaum, India in 2017 and has more than 21+ years of teaching experience. His interests include data mining, AI, and big data. He can be contacted at email: raghav.trg@gmail.com.

Vinith Rejathalal is an assistant professor (senior grade) in the Department of Artificial Intelligence at Amrita Vishwa Vidyapeetham, Coimbatore, India. He earned his Ph.D. from the National Institute of Technology Calicut in 2018. His research interests include applied ML, graph data analytics, and the application of DL models to graph data across various domains. He has presented his work at various conferences and published in reputed journals. He can be contacted at email: r_vinith@cb.amrita.edu.