Biomedical Signal Processing and Control 119 (2026) 109994

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Risk-informed design of Software as a Medical Device through Natural Language Processing techniques

Alessio Luschi [a,b], Alessandro Zazzeri [a], Gabriele Cevenini [a], Ernesto Iadanza [a,c,*]

[a] Department of Medical Biotechnologies, University of Siena, Via Aldo Moro, 2, Siena, 53100, Italy
[b] Digital Health and Clinical Engineering Divisions, International Federation for Medical and Biological Engineering, Paris, France
[c] Council of Societies, International Federation for Medical and Biological Engineering, Paris, France

ARTICLE INFO

Keywords:
Software as a Medical Device
Risk-informed design
Health information technology
Natural Language Processing
Deep learning
Failure classification
Medical device design

ABSTRACT

The rapid evolution of Software as a Medical Device (SaMD) for diagnostic and therapeutic applications requires evidence-based design strategies to maximise the benefit-to-risk ratio. This study aims to provide manufacturers with a framework for reducing risks by design. We also propose a novel standard classification of software failures related to Health Information Technologies (HIT) using a Natural Language Processing (NLP) multinomial classifier, shaping the entire design process of SaMD in an evidence-based, risk-aware manner.

Adverse event reports (2022–2024) were extracted from the FDA MAUDE database. HIT reports were identified using a binomial NLP classifier from the authors. A preliminary taxonomy of failure modes was derived from the literature and refined using self-supervised learning. K-modes clustering was applied to generate a balanced sample of 1048 records, then manually labelled and used to fine-tune the final classifier. Model performance was assessed through 10-fold cross-validation.

The multinomial classifier achieved cross-validated accuracies between 74.29% and 83.81% with an F1-score up to 0.87 for dominant classes. It enables rapid identification of recurring issues, helping developers prioritise design improvements based on real-world risks. Nine failure categories were also identified. Underrepresented categories exhibited lower performance due to the limited availability of training data. This study demonstrates the feasibility of integrating deep learning-based failure classification into SaMD design workflows and proposes a standard classification for HIT-related software failures. By leveraging insights from historical data, manufacturers can proactively identify and mitigate potential hazards, thereby enhancing both patient safety and regulatory compliance. This proactive, data-driven approach supports the creation of safer and more reliable biomedical devices and digital health technologies.

1. Introduction

The integration of advanced digital technologies into healthcare has transformed diagnostic and therapeutic practices, enabling precision medicine, real-time patient monitoring, and adaptive therapeutic interventions [1]. Among these innovations, Software as a Medical Device (SaMD) has emerged as a critical component of biomedical systems, ranging from clinical decision support to imaging analysis and implantable device management. Unlike traditional hardware-based devices, SaMD must address unique challenges related to software lifecycle management, human engineering, cybersecurity, and interoperability with health information technologies (HIT). Current industry practices are heavily focused on a post-market reactive approach, which often results in delayed detection of safety issues, failing to anticipate failure modes arising from complex software ecosystems [2]. Regulatory frameworks, such as the European Union Medical Device Regulation (EU-MDR) [3], as well as the guidelines of the International Medical Device Regulators Forum (IMDRF) [4] and international standards [5,6], highlight the necessity of risk-informed design processes that incorporate real-world evidence (RWE) from the earliest stages of development. However, this approach is hindered by the requirement of a significant amount of real-world data (RWD) (observational data related to outcomes in real-world contexts) to conduct an exhaustive assessment and extract significant evidence [7]. Moreover, to date, there is a lack of consistency among countries, which makes comparisons and benchmarking difficult. The same medical equipment can be classified using different codes and nomenclatures from country to country [8,9], whereas there is no international standard classification of failure codes for medical devices (MDs) and SaMD [10,11].

* Corresponding author at: Department of Medical Biotechnologies, University of Siena, Via Aldo Moro, 2, Siena, 53100, Italy.
E-mail address: ernesto.iadanza@unisi.it (E. Iadanza).

https://doi.org/10.1016/j.bspc.2026.109994
Received 29 October 2025; Received in revised form 11 February 2026; Accepted 24 February 2026
Available online 24 February 2026
1746-8094/© 2026 Published by Elsevier Ltd.

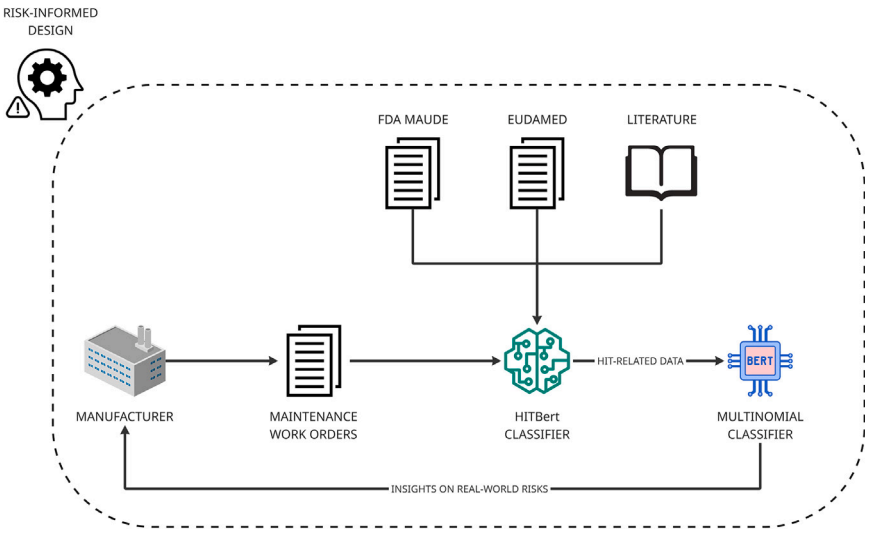


Fig. 1. The integration of the proposed framework within the developing process of SaMD for a risk-informed design.

Emerging methodologies based on Natural Language Processing (NLP) and deep learning present opportunities to leverage vast quantities of RWD from spontaneous reporting systems (SRSs) and electronic health records (EHRs) [12,13], potentially overcoming the highlighted limitations. Particularly, transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers), present robust methods for automating textual data analysis, significantly enhancing the capability to classify and analyse software-related device failures [14,15]. Applying these methods to classify and analyse HIT-related failures enables systematic identification of recurrent risk patterns, including software bugs, network errors, human factors, and design deficiencies [16,17]. A standardised classification of such failure modes would allow developers to incorporate risk insights directly into the SaMD design cycle, increasing reliability and reducing burden for manufacturers, thus improving both the design process and post-market surveillance activities.

The present study expands upon previous evidence-based maintenance approaches [18,19], proposing a framework that integrates within the designing process of SaMD, allowing for a risk-informed approach that permits prioritising improvements based on RWE (Fig. 1). Specifically, we introduce a deep learning-based framework that leverages NLP techniques to perform a multinomial classification of HIT-related failures extracted from the FDA MAUDE database. Classification is performed according to an innovative taxonomy, specifically designed for HIT-related failures. By linking failure patterns to root causes identifiable during the software lifecycle, the resulting framework provides manufacturers with actionable insights to anticipate and mitigate risks during software design and development phases, ultimately enhancing safety, regulatory compliance, and lifecycle performance of digital health technologies.

2. Literature review

A scientific review was initially carried out to develop an initial classification for SaMD failures using insights from existing literature. Verified high-level sources (open-access articles, reviews, and conference papers) related to software failure or malfunction in the medical subject areas, and selected between 2014 and 2024, were extracted from Scopus. Relevant articles were initially identified based on the relevance of titles and abstracts, and then through full-text analysis.

The literature analysis identified 17 full articles (Table 1). Building on these findings, literature-derived taxonomies of software failure modes were synthesised and aligned with design and development priorities for SaMD (Table 2).

3. Materials and methods

Fig. 2 illustrates the workflow of the proposed study. The process is structured into multiple linked stages, each aimed at progressively refining the previously obtained classification. Insights drawn from literature research and general domain knowledge guided the preliminary classification shown in Table 2. Afterwards, this initial framework was refined by comparing each class with tokens extracted from RWD processed using a BERT-based self-supervised classifier trained on HIT-related adverse event reports extracted from the MAUDE database between 2022 and 2024. This approach forced the model to infer and suggest the most probable cause of the software failure associated with each reported event by leveraging Masked Language Modelling (MLM). Lastly, an additional layer of polishing is exploited by reviewing the classification after the manual reading and labelling of a sample of reports obtained through K-mode clustering. This final subset was later exploited for fine-tuning a new BERT-based model to automatically perform a multinomial classification of reports within the identified HIT-related failure modes.

3.1. Data sources

Health agencies generally have two sorts of regulatory databases: SRS databases and recall/alert databases. The main publicly available SRS databases are:

- Manufacturer and User Facility Device Experience (MAUDE) by the US FDA Center for Devices & Radiological Health
- European Databank for Medical Devices (EUDAMED) by the EU Commission
- Database of Adverse Event Notifications (DAEN) by the Australian Therapeutic Goods Administration

Manufacturers are required to report incidents to these vigilance databases. Healthcare professionals, patients, and other organisations may report incidents at their discretion. The MAUDE database served as the primary source of RWD for this study, as it is the most widely used and publicly accessible SRS [36].

3.2. Data selection and preprocessing

In the MAUDE database, the data is structured across multiple tables, each dedicated to specific aspects of a report, such as patient details, device information, and other related attributes. This relational

Table 1
Selected articles after the literature review sorted by year.

Title	F. Author	Year	Ref
Enabling reliable usability assessment and comparative analysis of medical software: a comprehensive framework for multimodal biomedical imaging platforms	E. Denisova	2024	[20]
Insulin Pump-Associated Adverse Events in a Brazilian Reference Center for the Treatment of Diabetes Mellitus: Proposal for a Taxonomy of Device Failures in Adults, Adolescents, and Children	A. L. D. Neves	2024	[21]
Image-Guided Surgical Device Failures in Functional Endoscopic Sinus Surgery: A MAUDE Analysis	S. W. Hassanin	2023	[22]
Nurses' Perceptions about Smart Beds in Hospitals	S. H. Tak	2023	[23]
Software-related recalls in computer-assisted hip and knee arthroplasty	F. Castagnini	2023	[24]
Radiofrequency remote monitor software patch update without cybersecurity implantable cardioverter-defibrillator firmware update increases the risk of inappropriate implantable cardioverter-defibrillator therapies	X. Qian	2022	[25]
A model for the remote deployment, update, and safe recovery for commercial sensor-based IoT systems	A. Radovici	2020	[26]
Pacemaker firmware update and interrogation malfunction	J. Z. Lee	2019	[27]
Electrocardiogram failure in the operating room – bench testing to prevent bed-side disaster	B. Cowie	2018	[28]
Failure analysis for ultrasound machines in a radiology department after implementation of predictive maintenance method	G. Chu	2018	[29]
Toward safer health care: A review strategy of FDA medical device adverse event database to identify and categorize health information technology related events	H. Kang	2018	[17]
Adverse Events Involving Radiation Oncology Medical Devices: Comprehensive Analysis of US Food and Drug Administration Data, 1991 to 2015	M. J. Connor	2017	[30]
Error reporting from the da Vinci surgical system in robotic surgery: A Canadian multi-specialty experience at a single academic centre	E. Rajih	2017	[31]
Software-Related Recalls of Health Information Technology and Other Medical Devices: Implications for FDA Regulation of Digital Health	J. G. Ronquillo	2017	[32]
Analysis of clinical decision support system malfunctions: A case series and survey	A. Wright	2016	[33]
Malfunctions of robotic system in surgery: Role and responsibility of surgeon in legal point of view	A. Ferrarese	2016	[34]
Detecting software failures in the MAUDE database: A preliminary analysis	F. Pecoraro	2013	[35]

Table 2
Preliminary classification of HIT-related failure modes after literary research.

Class	Description	References
Perfective and Adaptive Maintenance	Problems occurring after the installation of a software update or a preventive maintenance event.	[24–27,30,33,34]
Corrective Maintenance	Software problems that arise spontaneously during device usage (bugs, code corruption, etc.).	[17,21,22,24,29–32,34,35]
Firmware	Problems occurring after a firmware update or in general, attributed to firmware (corruptions, bugs, etc.).	[25–27]
Cybersecurity	Problems related to security breaches (data theft, unauthorised access, etc.).	[25–27,32]
Data Corruption	Data loss not related to connection issues.	[17,21,33]
Connections and Environment	All problems related to data loss due to connection interruptions or absence, inability to connect two or more devices due to network issues.	[17,21,22,26,27,32,35]
User Error	Any problems caused by incorrect input or actions taken by human users.	[17,21,22,30,33,34]
GUI	All problems related to a non-user-friendly or misleading graphical user interface.	[20]

design optimises storage by avoiding redundancy, as shared attributes are stored only once and referenced as needed. Report keys link the tables together, ensuring accurate associations between records and their attributes for reliable data retrieval.

Only data from recent reports (2022–2024) were selected. This time frame was thought to be sufficient to keep the dataset at a manageable size for both storage and computational efficiency, while still capturing the most recent technological trends and reporting practices. Restricting the dataset to recent years was essential to ensure computational viability and focus on current technology developments, as SaMD adoption was far less widespread 10–15 years ago, resulting in significantly fewer reported cases. However, this decision may impact the dataset representativeness, as over the recent years, SaMD architecture, interoperability, and regulations have significantly changed. As a result, the current research may underrepresent failure modes seen in previous periods, particularly those associated with legacy systems and outdated communication protocols or regulatory frameworks.

The selection process produced a total of about 20 million entries, which were later imported into Microsoft SQL Server and preprocessed to ensure dataset integrity by removing duplicates and records with useless information. The process significantly reduced the dataset to a more manageable subset of about 6 million unique entries.

The initial filtering of HIT-related adverse event reports was performed using a previously developed and validated binary text classifier (hereafter referred to as HITBert). The model is based on a fine-tuned ClinicalBERT architecture trained on 3705 adverse event reports extracted from the MAUDE database, manually labelled by domain experts as HIT or non-HIT. Model performance was evaluated using a 10-fold cross-validation strategy and an independent held-out test set, achieving accuracy, precision, recall, and F1-score values exceeding 0.96, with recall values close to 1.0 for the HIT class. Explainable AI techniques (SHapley Additive exPlanations — SHAP [37] and LIME [38]) were applied to assess feature relevance and decision consistency, confirming that the model relies on clinically and technically meaningful terms [19]. The use of a high-recall binary classifier

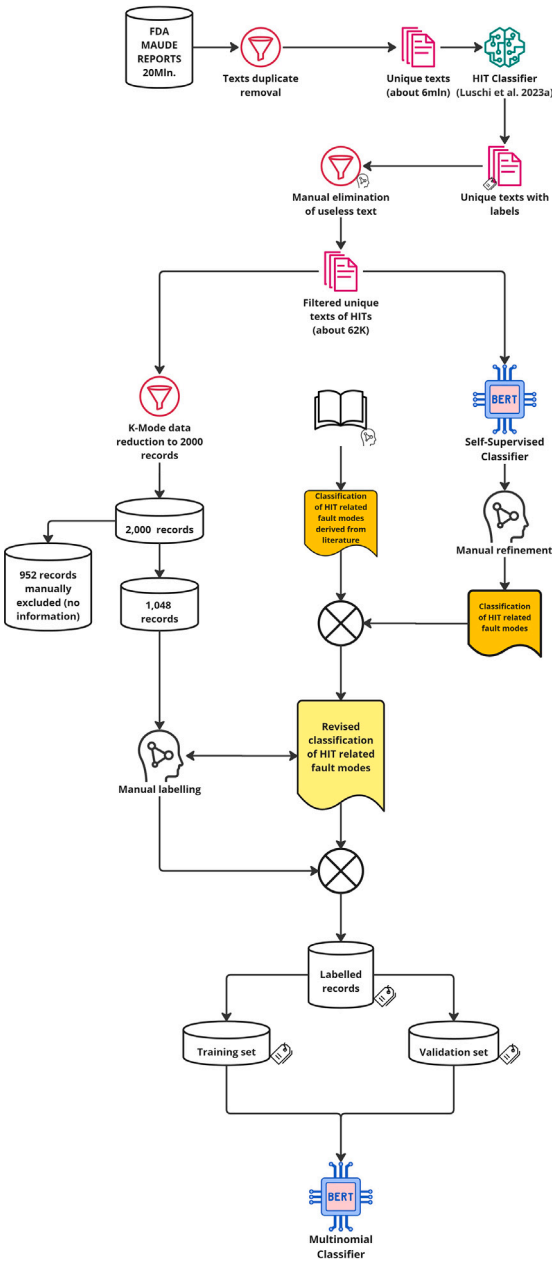


Fig. 2. Study workflow.

at this stage ensures that the downstream stages are applied only to reports with a high likelihood of being HIT-related, reducing noise while minimising false exclusions.

The HITBert model was initially leveraged to extract only records about HIT-related adverse events from the obtained dataset of 6 million unique entries. After the classification, several data inspections were carried out to verify consistency with the expected results. These inspections revealed some records containing texts irrelevant to the study, such as entries lacking failure descriptions or internal communications (e.g., failure code updates, manufacturer name corrections, or error reports). Although limited in number and not particularly problematic, these issues were mitigated through partial manual filtering. Specifically, the dataset was sorted by text length (shorter entries were more often irrelevant), and these were reviewed and removed until the issue was no longer significant. The final number of filtered HIT-related records was about 62,000.

3.2.1. Data quality and reliability considerations

Public post-market surveillance databases, such as the MAUDE, are known to contain substantial noise, missing information, underreporting, and occasional inconsistencies across reports. These characteristics may affect the reliability of downstream analyses if not explicitly addressed. To mitigate these issues, a data quality–oriented preprocessing strategy was adopted before taxonomy development and model training.

First, only reports containing a sufficiently informative narrative description were retained, as free-text fields constitute the primary source of evidence for failure classification in this study. Records with missing, extremely short, or non-informative textual descriptions were excluded through listwise deletion. Structured variables exhibiting high rates of missingness or inconsistent usage were not used as primary features, thereby avoiding bias introduced by unreliable metadata.

Second, the use of an initial high-recall HITBert classifier [19] ensured that only reports with a high likelihood of being HIT-related were propagated to subsequent stages, reducing noise introduced by unrelated adverse events. Manual expert review of the clustered subset further contributed to filtering ambiguous or contradictory cases during taxonomy refinement.

While underreporting and reporting bias remain inherent limitations of SRSs, the adopted quality-control strategy aimed to maximise internal consistency and semantic reliability of the analysed corpus. Following this filtering and validation process, MAUDE remains the most comprehensive and practical data source currently available for large-scale, real-world analysis of failures related to HIT and SaMD, particularly for studies focused on post-market risk patterns rather than incidence estimation.

3.3. Self-supervised learning

Self-supervised learning (SSL) was employed to automatically extract tokens related to possible failure modes from the identified HIT-related adverse event reports with zero-shot classification. BERT-based models were leveraged since they are pretrained with MLM to predict masked tokens in text. MLM is the task of masking some of the words in a sentence and predicting which words should replace those masks [15], and can be used to train large NLP models for domain-specific problems, such as in this case. A pretrained BERT model was exploited to classify reports by providing the masked ending sentence in the prompt: “the failure of this medical device is caused by a [MASK] error”. Five different pre-trained models (bert_base_uncased, bert_large_uncased, roberta_base, roberta_large and Bio_ClinicalBERT) were tested during preliminary experiments. Model performance was assessed qualitatively rather than through standard quantitative metrics. Specifically, candidate models were evaluated by analysing the semantic relevance and interpretability of the output tokens generated by MLM, comparing them against expert judgement derived from manual inspection of the corresponding adverse event reports. This qualitative evaluation approach was selected because the objective of this phase was not predictive accuracy, but the identification of meaningful failure-related concepts to guide taxonomy refinement. Consequently, numerical performance metrics were not collected, as they would not have directly reflected the quality or clinical relevance of the extracted tokens. This qualitative comparison was limited to taxonomy refinement and did not influence the subsequent supervised training or evaluation of the multinomial classifier, which was assessed independently using standard quantitative metrics. Therefore, the absence of numerical benchmarks at this exploratory stage does not affect the validity or interpretability of the final classifier performance results. The bert_base_uncased model reflected a trade-off between semantic clarity, interpretability of generated tokens, and robustness across heterogeneous reporting styles present in MAUDE narratives (Table 3). The top-20 most frequent tokens identified by the chosen model were

Table 3
Qualitative comparison of pre-trained transformer models used during exploratory evaluation.

Model	Domain adaptation	Token semantic relevance	Consistency with expert interpretation
bert_base_uncased	General	High	High
bert_large_uncased	General	High	Moderate
roberta_base	General	Moderate	Moderate
roberta_large	General	Moderate	Moderate
Bio_ClinicalBERT	Clinical	Low–Moderate	Low

then exploited to perform a revision of the preliminary classification obtained after the literature review.

The development of the failure taxonomy followed a hierarchical and evidence-informed integration strategy. Categories identified through the literature review (Table 2) were treated as the conceptual baseline, ensuring consistency with established failure models and regulatory discourse. MLM with the chosen bert_base_uncased model was then applied to adverse event reports to extract empirically salient tokens reflecting how failures are described in real-world reports.

When MLM-inferred tokens aligned with literature-derived categories, they were used to refine class names and descriptions to improve semantic clarity. When MLM outputs suggested concepts not explicitly represented in the literature, their inclusion was contingent on frequency, semantic coherence across reports, and expert validation. Conversely, literature-derived categories that were weakly represented in MLM outputs (e.g., Cybersecurity) were retained to preserve conceptual completeness, as their absence in token frequencies reflects rarity rather than irrelevance. Conflicts between evidence sources were resolved through expert consensus, prioritising clinical relevance, regulatory significance, and interpretability.

3.4. K-mode clustering

In parallel, manual labelling of the filtered HIT-related reports was requested for fine-tuning the final BERT model for multinomial classification (Fig. 2). As this operation was practically unfeasible due to the huge investment of time and human resources needed, a subset of data was deemed necessary to be identified for annotation. Following an evaluation of the general length and quality of the records, the sample size was set at no more than 2000 entries, balancing feasibility with representativeness. The major key challenge was to devise a reduction method that preserved the overall distribution and informational content of the dataset. A review of data reduction methodologies indicated that K-means clustering is frequently employed for sampling unbalanced datasets [39]. However, given that the attributes in this dataset were predominantly categorical, the direct application of K-means was not viable, as computing means on non-numeric data is not meaningful. Consequently, class imbalance was addressed primarily through K-modes clustering, which substitutes the calculation of the mean with the mode [40]. The algorithm was used to construct a reduced yet information-rich subset while preserving the heterogeneity of device types, manufacturers, clinical fields, and textual descriptions. This approach was selected to maximise semantic diversity and prevent dominance of highly frequent failure categories during manual annotation. Alternative imbalance mitigation techniques, such as oversampling or synthetic data generation (e.g., Synthetic Minority Over-sampling Technique — SMOTE), were not applied in this phase. While effective in structured feature spaces, their application to free-text clinical narratives requires operating on dense embedding representations and may introduce synthetic samples with limited clinical plausibility if not carefully constrained [41–43]. Given the safety-critical context of medical device failure analysis, preserving fidelity to real-world reports was prioritised.

The K-modes clustering enabled the grouping of records with similar characteristics into K clusters, from which proportional samples were drawn. Such an approach facilitated the construction of a balanced and representative training dataset, thereby maximising the informational value for subsequent model development. The K-modes algorithm begins by randomly selecting K elements from the dataset to serve as initial cluster centres. The remaining data points are then assigned to the cluster whose centre is closest to them. Once all points have been assigned and the clusters formed, the mode of each attribute within each cluster is calculated, and this set of modes becomes the new cluster centre. This process is repeated iteratively until the algorithm converges, meaning no data points change clusters between iterations. Since convergence is not guaranteed, a maximum of 10 iterations was set. Due to the categorical nature of the attributes, the most suitable metric to compute distances between records was the Hamming distance [44], which measures dissimilarity by assigning a value of +1 for each attribute that differs between two records:

$$H_{x,y} = \sum_{i=1}^{n} \begin{cases} 1 & \text{if } x_i \neq y_i \\ 0 & \text{if } x_i = y_i \end{cases} \tag{1}$$

where x_i is the value of attribute i for the data x, y_i is the value of attribute i for the data y, and N is the number of attributes.

The number of clusters, K, is not fixed and is typically determined by evaluating how well the distance metrics are optimised within the resulting clusters. The optimal number of clusters was identified through an iterative process. Multiple runs of the K-modes algorithm were performed, each time increasing the number of clusters. For each configuration, the average intra- and inter-cluster distances were calculated, and their difference was analysed to identify the best clustering solution. Once the optimal value of K was determined, the algorithm was executed again, and clusters were sampled by selecting the N records closest to each cluster centre, with N defined as:

$$N = \frac{2000}{K} \tag{2}$$

3.5. Multinomial classification

The final phase of the study focused on the multinomial classification of records using the final designed taxonomy. The taxonomy validation relied on expert consensus rather than purely statistical inter-rater metrics. While coefficients such as Cohen's κ are useful when class definitions are fixed a priori, they may underestimate agreement during iterative taxonomy construction where categories are progressively refined [45,46]. In this context, consensus-based expert validation was prioritised to ensure clinical relevance, semantic clarity, and alignment with real-world failure reporting practices. Two domain experts with backgrounds in clinical engineering and health information technology (A.L. and E.I.) independently reviewed and labelled the clustered subset of HIT-related adverse event reports, assigning each to the most appropriate failure category while discarding irrelevant entries. Discrepancies in category assignment were discussed jointly and resolved through consensus, leading to iterative clarification and refinement of class definitions.

The labelled data was then used to fine-tune a pre-trained BERT model. The dataset was split into training and validation sets following an 80:20 ratio. Performance was assessed using standard metrics for multiclass classification: accuracy, precision, recall, and F1-score. After hyperparameter tuning, a 10-fold cross-validation was performed to ensure robustness and reduce the risk of bias introduced by dataset splitting.

4. Results

The first proposed taxonomy of failure modes related to SaMD is shown in Table 2. The eight classes emerged from the literature analysis performed as the first step of the study.

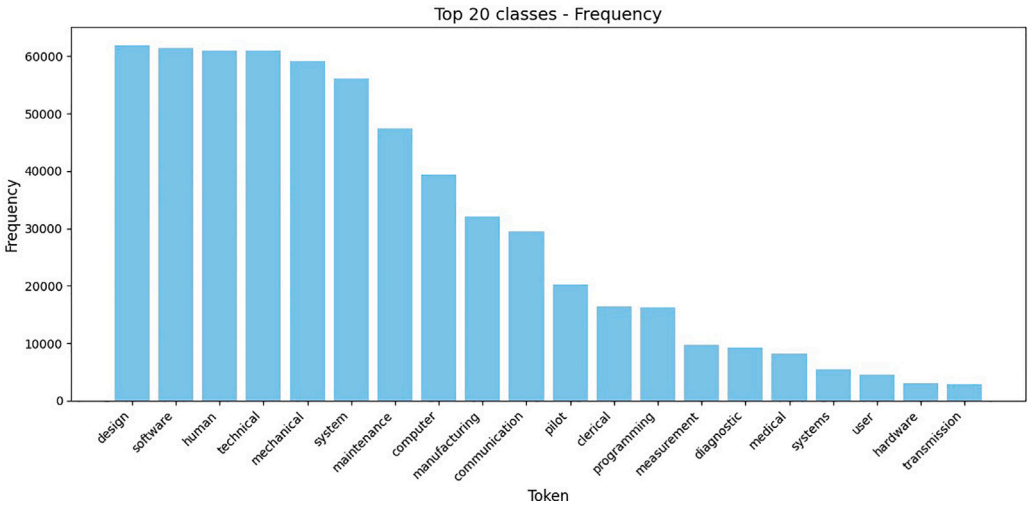


Fig. 3. Top 20 tokens in frequency order that emerged from self-supervised learning.

Table 4
Selected tokens and relative frequencies after the self-supervised learning analysis.

Token	Frequency
Design	61.92
Software	6.47
Human	60.99
Maintenance	47.37
Computer	39.33
Manufacturing	32.07
Communication	29.45
User	4.56
Network	2.46
Driver	1.20
Database	63.00
Data	25.00

4.1. Self-supervised learning

Fig. 3 shows the first 20 most frequent tokens that were identified by applying MLM and self-supervised learning to the subset of preprocessed unique 62,000 HIT-related records extracted from the MAUDE database and filtered by the HITBert model [19].

The results align well with the provided prompt sentence and, more importantly, with the intended objective. The only notable limitation is the occasional occurrence of tokens unrelated to HITs. This outcome is expected, given that making the prompt sentence more specific would have negatively affected overall performance, considering the inherent imprecision of a general-purpose model without fine-tuning. After analysing all the proposed tokens, those most consistent with the intended objective are listed in Table 4.

By analysing the extracted tokens, the initial classification was revised, trying to match the evidence that emerged from the identification of tokens related to possible software failures to the initial proposed categories. This revision process ensured that the taxonomy remained grounded in established knowledge while being empirically adapted to reflect dominant and emerging failure patterns observed in post-market surveillance data. The new classification is listed in Table 5.

This revised classification introduces updated names for classes and more concise and meaningful descriptions. *Cybersecurity* class remains from the first classification, even though no associated token was identified during the self-supervised learning. This decision reflects the prioritisation of regulatory relevance and risk severity over empirical frequency, as it retains evidence about a failure mode that cannot be discarded (according to the authors' expertise), representing high-impact hazards that cannot be excluded from a comprehensive SaMD failure taxonomy.

4.2. K-mode clustering

During the K-mode cluster optimisation procedure, the 62,000 identified samples were processed at each iteration. To balance computational cost and result quality, the number of clusters to be tested was varied between 0 and 1000, using increments of 100. Fig. 4 shows that the optimal configuration corresponds to $K = 1000$ clusters, which minimises the average intra-cluster distance without negatively affecting the average inter-cluster distance (Eq. (1)).

Thus, according to Eq. (2), clusters were sampled by selecting the two closest records to each cluster centre ($N = 2$). Since maximising the variety of cases present in the samples was the aim, seven MAUDE attributes were used, deemed the most relevant:

- Year of the event (YEAR)
- Type of event (EVENT_TYPE)
- Name of the manufacturer (MANUFACTURER_D_NAME)
- Brand of the device (BRAND_NAME)
- Generic name of the device (GENERIC_NAME)
- Medical field (MEDICAL_FIELD)
- Report text (FOI_TEXT)

Afterwards, 952 out of 2000 records were manually excluded because they contained no useful textual information, resulting in a final subset of 1048 records. These records were ultimately labelled according to the classification proposed in Table 5 to provide an annotated dataset for fine-tuning the final BERT-based multinomial classifier. The procedure revealed that the proposed classes were already highly consistent with RWE; the only required adjustment was the addition of an extra class to account for calibration failure events, resulting in the final version of the proposed classification standard for HIT-related software failures (Table 6).

4.3. Statistical analysis

Before proceeding to the final multinomial classification with NLP, a descriptive statistical analysis was conducted on the annotated dataset of 1048 records. Fig. 5 shows the number of HIT-related adverse event reports grouped by year for both the annotated subset and the full dataset, while Fig. 6 shows the distribution of reports in the identified classes year-wise.

Table 5
New revised classification of HIT-related failures after self-supervised learning.

Class	Description	Associated token
Induced by Service (SIF)	All problems occurring during or after maintenance was performed.	Maintenance
Software	All software problems that arise spontaneously during device usage (bugs, code corruption, etc.).	Software, computer
Driver	Problems occurring after a driver update or in general, attributed to drivers (corruptions, bugs, etc.).	Driver
Cybersecurity	Problems related to security breaches (data theft, unauthorised access, etc.).	–
Data Quality	Data loss not related to connection issues.	Database, data
Network	Problems related to data loss due to connection interruptions or absence, inability to connect two or more devices due to network issues.	Communication, network
Human Factors	Problems caused by incorrect input or user actions, either by accident or due to a non-user-friendly GUI.	Human, user
Design and Manufacturing	Problems related to defects in software coding.	Design, manufacturing

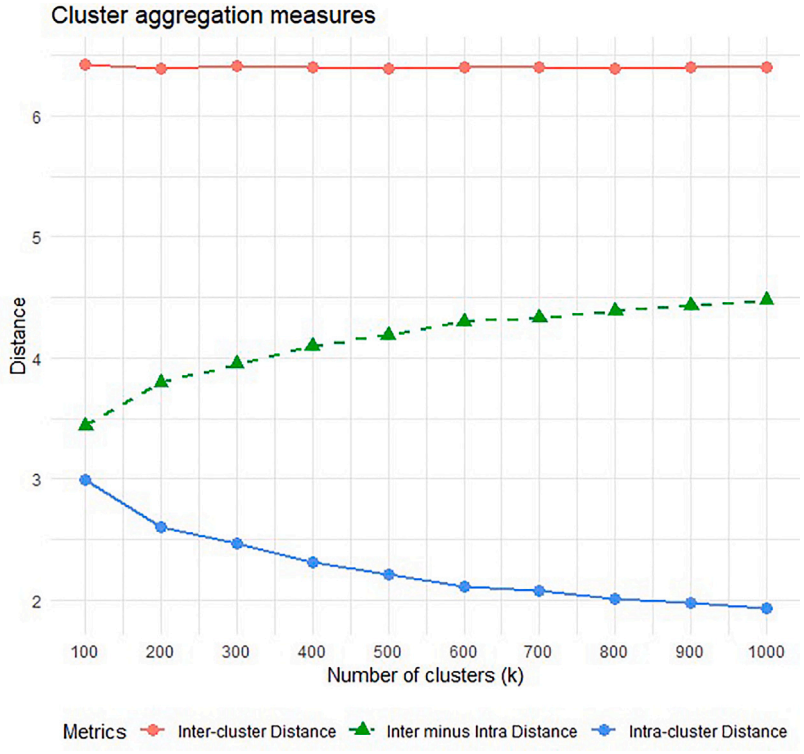


Fig. 4. K-mode results with varying cluster numbers.

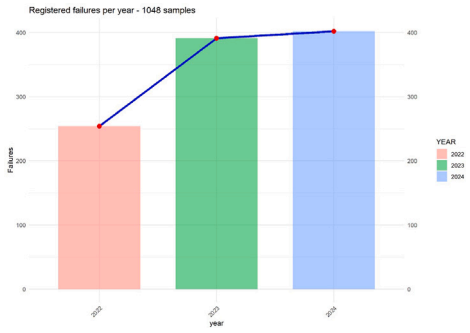
Table 6
Final proposed classification for HIT-related software failures.

Class	Description
Service-Induced Failure (SIF)	All problems occurring during or after maintenance was performed.
Software	All software problems that arise spontaneously during device usage (bugs, code corruption, etc.).
Driver	Problems occurring after a driver update or in general, attributed to drivers (corruptions, bugs, etc.).
Cybersecurity	Problems related to security breaches (data theft, unauthorised access, etc.).
Data Quality	Data loss not related to connection issues.
Network (NET)	Problems related to data loss due to connection interruptions or absence, inability to connect two or more devices due to network issues.
Human Factors	Problems caused by incorrect input or user actions, either by accident or due to a non-user-friendly GUI.
Design and Manufacturing	Problems related to defects in software coding.
Performance and Calibration	Problems related to precision, accuracy, and calibration.

Fig. 7 highlights the distribution of the top-10 frequent records for the most numerous classes (*Network* and *Software*) across four MAUDE attributes: brand of the device (BRAND_NAME), generic name of the device (GENERIC_NAME), name of the manufacturer (MANUFACTURER_D_NAME), and medical field (MEDICAL_FIELD).

4.4. Multinomial classification

Various experiments were initially conducted to tweak the model parameters. Dropout hyperparameters have been constantly set to 0.5 for the attention layer and 0.1 for the hidden layer [47]. Five different



(a) Registered HIT-related failures per year on 1048 records subset.

(b) Registered HIT-related failures per year on the 118,000 filtered HIT-related adverse event reports, before K-mode clustering.

Fig. 5. Comparison between the distribution of samples over years in the labelled dataset 5(a) and the whole dataset 5(b).

models (roberta_base, roberta_large, Bio_ClinicalBERT, bert_base_uncased, and bert_large_uncased) with three different activation functions — the Sigmoid Linear Unit (SiLU), the Rectifier Linear Unit (ReLU), and the Gaussian Error Linear Unit (GELU), three learning rates for the optimisation algorithm ($2e^{-5}$, $3e^{-5}$, and $5e^{-5}$), and three batch sizes (8, 16, and 32) were tested as suggested by Devlin et al. [15]. Tests were also conducted on the number of frozen layers to achieve the best performance: 0, 4, 8, or 12 encoder layers were independently frozen. The best performances were achieved with the bert_base_uncased model with 4 frozen layers, the GELU activation function, a batch size of 16, and $5e^{-5}$ learning rate. The model was trained for a total of 15 epochs.

Fig. 8 highlights how the model began to overfit after the first seven epochs. The observed trend aligns with the overall technique of fine-tuning BERT-based models for only a few epochs [15]. Thus, the model was trained for only seven epochs to prevent overfitting, obtaining 0.7905 general accuracy, 0.6376 precision, 0.5449 recall, and 0.5291 F1-score (Table 7). The confusion matrix is shown in Fig. 9. *Cybersecurity* and *SIF* classes had not enough data to perform any training or validation, so they were excluded from the performance analysis.

Finally, a brand new model was trained on the same dataset of 1048 records with the identified tuned hyperparameters and k-fold cross-validation (10 folds). Results are shown in Tables 8–10.

5. Discussion

Fig. 1 highlights how the proposed framework integrates into the SaMD design process, allowing for a risk-informed approach that prioritises improvements based on RWE.

The proposed multinomial classifier, together with the previous HITBert classifier [19], and the identified taxonomy of software-related failure modes, allows a rapid identification of recurring issues that emerge from real-world insights, enabling manufacturers to proactively mitigate potential hazards and improve outcomes. Overall model performance exhibited substantial variability across classes. As shown in Table 8, classes with higher representation in the training data, such as *Software* and *Network*, achieved consistently higher accuracy scores across folds. Tables 9 and 10 also show that *Software* and *Network* exhibit higher mean accuracy and F1-scores with relatively low standard deviation, indicating consistent performance across folds. Class-specific metrics (Table 8) show that lower and more unstable performance, with F1-scores below 0.6, predominantly occur in underrepresented classes with limited training samples. These results reflect the strong influence of class imbalance on classification outcomes and highlight that categories with sufficient sample sizes can achieve performance consistent with a well-performing classifier capable of correctly categorising their entries. The validation confusion matrix (Fig. 9) provides additional insight into these results. Correct predictions are strongly concentrated along the diagonal for the most frequent classes, indicating reliable discrimination when sufficient training examples are available. Conversely, misclassifications primarily affect underrepresented categories and tend to occur toward semantically related classes, rather than being uniformly distributed. This pattern suggests that reduced F1-scores are largely driven by data scarcity and class overlap, rather than by systematic model failure. Hence, the proposed multinomial classifier shows robust performance for dominant and recurrent failure categories, which are most relevant for risk-informed design prioritisation. However, performance for rare failure modes remains limited and should be interpreted cautiously, particularly in safety-critical contexts. Specifically, classifications involving underrepresented categories, such as *Human Factors* or *Design and Manufacturing*, should currently be considered as decision-support outputs requiring continued expert review rather than fully automated determinations, until larger and more balanced datasets become available.

The training data can be further expanded to include also adverse event reports originating from different SRSs (i.e., EUDAMED or DAEN), maintenance work orders provided directly from manufacturers, and actionable insights from academic literature mined from databases such as PubMed and Scopus.

The final proposed standard classification for HIT-related failures (Table 6) is considered to cover all possible failure modes that may be related to SaMD, as it incorporates evidence coming from different sources (scientific literature, outcomes of a self-supervised learning NLP algorithm applied on records extracted from the MAUDE, and manual refining done by experts). The slight edits that differentiate the final version from the previous ones (Tables 2, 5, 6) highlight how each method leveraged during the process produced overlapping results, reflecting a global significance of the identified classes. Unlike prior works primarily focused on post-market maintenance optimisation [18, 19], this study repositions failure analysis as a design-enabling tool, directly informing early-stage software engineering and risk management processes. The standardised taxonomy of nine HIT-related failure classes offers actionable insights for manufacturers seeking to align with regulatory frameworks and guidelines, which emphasise lifecycle safety and usability [3,4]. By identifying dominant failure categories, such as network and software errors, developers can incorporate design mitigations (e.g., redundancy in network communication, robust error handling) during the architecture and coding phases. This proactive use of failure intelligence supports ISO 14971 [5], and IEC 62304 [6] processes, bridging post-market evidence with pre-market verification and validation activities.

Table 11 explicitly maps taxonomy outputs to key phases of the risk management lifecycle defined in ISO 14971 and IMDRF guidance to clarify how the proposed taxonomy supports regulatory risk management activities. The taxonomy is intended as an operational tool that

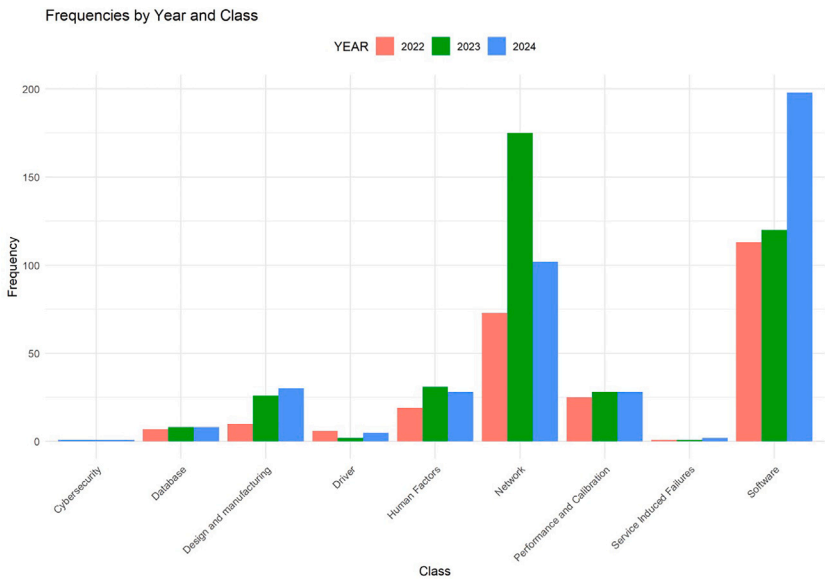


Fig. 6. Frequencies of HIT-related failures grouped by year and class.

Table 7
Class-specific metrics of single fold validation of the BERT-based classifier. *Cybersecurity* and *SIF* classes are excluded due to a lack of data.

Class	Accuracy	Precision	Recall	F1-score
Software	0.8906	0.8667	0.8966	0.8814
Driver	0.0000	0.0000	0.0000	0.0000
Database	0.2000	1.0000	0.2000	0.3333
Network	0.8571	0.9375	0.8571	0.8955
Human Factors	0.2500	0.5714	0.2500	0.3478
Design and Manufacturing	0.9231	0.4000	0.9231	0.5581
Performance and Calibration	0.6875	0.6875	0.6875	0.6875

supports hazard identification, verification planning, and validation of SaMD systems by grounding these activities in empirically observed post-market failure patterns.

Examining the class composition, a first observation arises from comparing the temporal distribution of events in the sampled subset (Fig. 5). The analysis reveals an upward trend in failure events over time, consistent with the growing technological adoption in the field. The close similarity between the two distributions further supports the representativeness of the sampled dataset obtained with K-mode clustering relative to the full dataset. Fig. 6 illustrates the variation in report frequencies over time for each class, highlighting both the previously mentioned issue of under-sampled classes and its extent. Beyond this, some noteworthy patterns emerge, such as the general increase in reported failures over the years, which aligns with expectations. The most significant deviation is observed in the *Network* class, which exhibits a marked peak in 2023. To investigate this anomaly, a focused analysis of the class was undertaken, identifying the ten most problematic devices for each of the three years examined. Fig. 10 indicates that the abnormal behaviour observed in this class is likely attributable to a single device. Across all three years analysed, the most frequently malfunctioning device was the *Freestyle Libre 2*, a glucose monitoring device for diabetic patients.

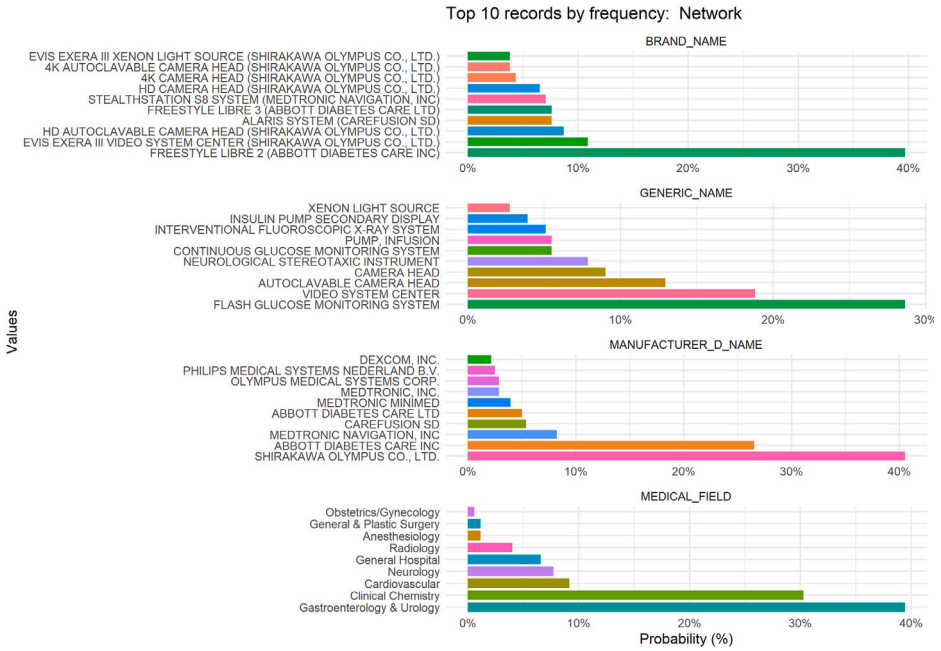
A closer examination of the relative frequencies among the top-10 devices reveals a pronounced increase in failures for the specific MD in 2023, during which it accounted for half of the reported failures within the top ten. In 2024, however, the failure rate returned to a more acceptable level. This pattern suggests that the anomalous peak observed in 2023 may have resulted from a widely reported product flaw during that year, which subsequently led to an increase in registered events. Once the issue was resolved by the manufacturer, malfunction reports declined in 2024.

Fig. 7(b) highlights that the majority of failures for the *Software* class happen in radiological, cardiovascular, and neurological applications, which is coherent with the high number of SaMD that are commonly utilised in these fields, ranging from CT scans, MRI, angiography equipment, reporting workstations, RIS (Radiological Information System), CIS (Cardiological Information System), and PACS (Picture Archiving and Communication System). Notably, an interventional fluoroscopy X-ray system, a medical imaging device that uses real-time X-ray technology to guide minimally invasive procedures, is the type of device that led to the majority of reported adverse events for this class.

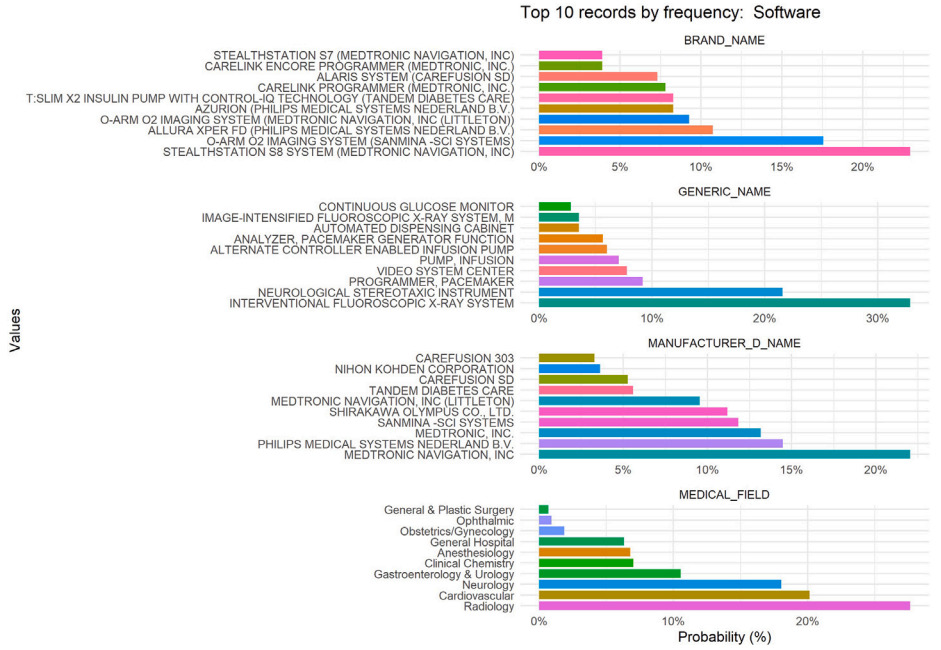
5.1. Practical application examples

The following examples illustrate how the proposed taxonomy can function as a decision-support tool within a risk management framework, enabling manufacturers to translate post-market evidence into concrete design, verification, and validation actions.

- **Design refinement driven by recurrent Software failures**: In the first scenario, a manufacturer is developing a network-enabled SaMD that performs automated data acquisition and decision support. The application of the proposed taxonomy to post-market surveillance data reveals a high prevalence of *Software* and *Network* failure categories, consistent with the dominant classes identified in this study. Based on these findings, the manufacturer may prioritise design changes such as improved exception handling for data transmission errors, enhanced logging mechanisms, and redundancy in network communication modules. In parallel, verification activities can be adjusted to include stress testing under degraded connectivity conditions, directly targeting failure patterns observed in real-world deployments.



(a) Distribution of the top-10 records by frequency for the class *Network* across four MAUDE attributes.



(b) Distribution of the top-10 records by frequency for the class *Software* across four MAUDE attributes.

Fig. 7. Comparison between the distribution of the top-10 records by frequency for the classes *Network* 7(a) and *Software* 7(b) across four MAUDE attributes.

- **Validation planning informed by Human Factors failures**: In a second scenario, a manufacturer applies the taxonomy during validation of a clinical decision support system intended for use in high-pressure clinical environments. The taxonomy highlights a non-negligible proportion of *Human Factors*-related failures, including user interaction errors and workflow mismatches. In response, the manufacturer may expand validation protocols to include simulated clinical workflows, usability testing with representative end users, and evaluation of interface clarity under time-constrained conditions. By explicitly linking validation activities to empirically observed failure categories, the taxonomy supports more realistic and risk-informed validation strategies.

5.2. Limitations

5.2.1. Dataset and reporting bias

A key limitation of this study arises from the exclusive use of the MAUDE database as the source of RWD. MAUDE primarily reflects adverse event reporting practices within the US and may underrepresent

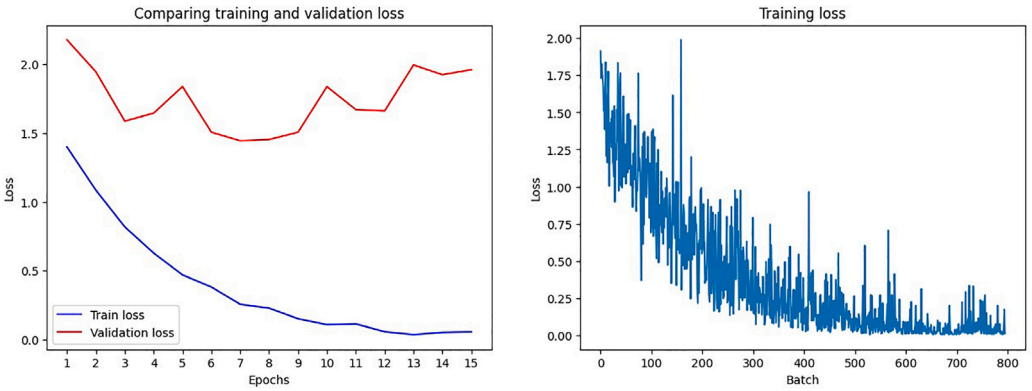


Fig. 8. Training and validation loss through the epochs.

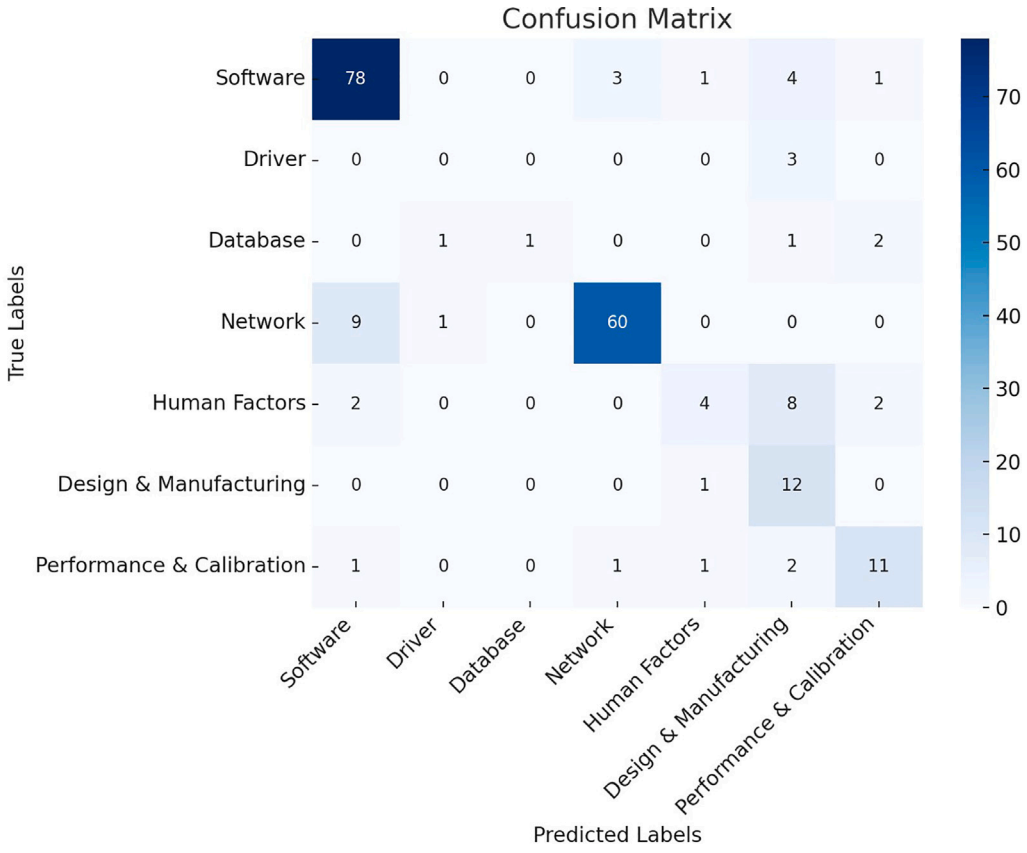


Fig. 9. Validation confusion matrix.

devices marketed predominantly outside the country, as well as region-specific regulatory, organisational, or clinical workflows. In addition, SRSs are known to exhibit reporting biases, including underreporting, variability in report quality, and temporal surges associated with recalls or regulatory actions. Consequently, the observed distribution of failure categories may not fully reflect the global landscape of SaMD failures.

An additional limitation concerns the temporal scope of the data. Restricting the analysis to adverse event reports from 2022–2024 may limit the generalisability of the identified failure taxonomy across longer device lifecycles. Earlier generations of SaMD may exhibit different failure patterns that are not fully captured in the selected period. Moreover, MAUDE reporting practices and manufacturer awareness have evolved, potentially introducing reporting biases in the analysed reports that may have influenced the observed distribution of failure modes.

5.2.2. Cross-database generalisability

Although the proposed classifier was trained and validated using MAUDE data, its performance on other vigilance databases (e.g., the European EUDAMED or the Australian DAEN) cannot be assumed without further investigation. Differences in reporting structure, terminology, language usage, and regulatory context may affect model performance if applied without retraining. Nevertheless, the taxonomy itself is designed to be conceptually database-agnostic, capturing failure modes that are independent of specific regulatory implementations. Future work may evaluate zero-shot or limited fine-tuning transfer of the classifier across databases to assess robustness and generalisability in multinational surveillance settings.

5.2.3. Ontology alignment and semantic interoperability

The current taxonomy was developed as a data-driven classification tailored to post-market failure analysis and SaMD design support,

Table 8
Class-specific performance metrics for ten-fold cross-validation.

Fold	Class	Accuracy	Precision	Recall	F1-score	Specificity
1	Software	0.8864	0.7647	0.8864	0.8211	0.8033
	Driver	0.0000	0.0000	0.0000	0.0000	1.0000
	Database	0.0000	0.0000	0.0000	0.0000	1.0000
	Network	0.8857	0.8378	0.8857	0.8611	0.9143
	Human Factors	0.6250	1.0000	0.6250	0.7692	1.0000
	Design and Manufacturing	0.7143	1.0000	0.7143	0.8333	1.0000
	Performance and Calibration	0.6250	0.7143	0.6250	0.6667	0.9794
2	Software	0.9091	0.8333	0.9091	0.8696	0.8689
	Driver	0.0000	0.0000	0.0000	0.0000	1.0000
	Database	0.5000	1.0000	0.5000	0.6667	1.0000
	Network	0.8571	0.9677	0.8571	0.9091	0.9857
	Human Factors	0.1250	0.5000	0.1250	0.2000	0.9897
	Design and Manufacturing	0.5000	0.3333	0.5000	0.4000	0.9394
	Performance and Calibration	0.8750	0.5000	0.8750	0.6364	0.9278
3	Software	0.8409	0.7115	0.8409	0.7708	0.7541
	Driver	0.5000	1.0000	0.5000	0.6667	1.0000
	Database	0.5000	0.5000	0.5000	0.5000	0.9903
	Network	1.0000	0.9211	1.0000	0.9589	0.9571
	Human Factors	0.0000	0.0000	0.0000	0.0000	0.9897
	Design and Manufacturing	0.5000	0.6000	0.5000	0.5455	0.9798
	Performance and Calibration	0.5000	0.6667	0.5000	0.5714	0.9794
4	Software	0.8636	0.7917	0.8636	0.8261	0.8361
	Driver	0.0000	0.0000	0.0000	0.0000	0.9903
	Database	0.0000	0.0000	0.0000	0.0000	0.9804
	Network	0.8286	0.8286	0.8286	0.8286	0.9143
	Human Factors	0.2857	0.4000	0.2857	0.3333	0.9694
	Design and Manufacturing	0.6667	0.6667	0.6667	0.6667	0.9798
	Performance and Calibration	0.6250	0.6250	0.6250	0.6250	0.9691
5	Software	0.8837	0.9268	0.8837	0.9048	0.9516
	Driver	0.0000	0.0000	0.0000	0.0000	1.0000
	Database	1.0000	0.7500	1.0000	0.8571	0.9902
	Network	0.9429	0.7857	0.9429	0.8571	0.8714
	Human Factors	0.6250	0.5556	0.6250	0.5882	0.9588
	Design and Manufacturing	0.5714	1.0000	0.5714	0.7273	1.0000
	Performance and Calibration	0.6250	1.0000	0.6250	0.7692	1.0000
6	Software	0.8140	0.8537	0.8140	0.8333	0.9032
	Driver	0.0000	0.0000	0.0000	0.0000	1.0000
	Database	0.3333	0.5000	0.3333	0.4000	0.9902
	Network	0.8857	0.8857	0.8857	0.8857	0.9429
	Human Factors	0.7500	0.4615	0.7500	0.5714	0.9278
	Design and Manufacturing	0.2857	1.0000	0.2857	0.4444	1.0000
	Performance and Calibration	0.7500	1.0000	0.7500	0.8571	1.0000
7	Software	0.8372	0.8571	0.8372	0.8471	0.9032
	Driver	0.0000	0.0000	0.0000	0.0000	1.0000
	Database	0.0000	0.0000	0.0000	0.0000	1.0000
	Network	0.9714	0.8095	0.9714	0.8831	0.8857
	Human Factors	0.2500	0.3333	0.2500	0.2857	0.9588
	Design and Manufacturing	0.4286	0.3750	0.4286	0.4000	0.9490
	Performance and Calibration	0.6667	0.8571	0.6667	0.7500	0.9896
8	Software	0.8605	0.8222	0.8605	0.8409	0.8689
	Driver	0.0000	0.0000	0.0000	0.0000	0.9903
	Database	1.0000	1.0000	1.0000	1.0000	1.0000
	Network	0.8286	0.8056	0.8286	0.8169	0.8986
	Human Factors	0.3750	0.7500	0.3750	0.5000	0.9896
	Design and Manufacturing	0.7143	0.7143	0.7143	0.7143	0.9794
	Performance and Calibration	0.7500	0.6667	0.7500	0.7059	0.9688
9	Software	0.8837	0.7170	0.8837	0.7917	0.7541
	Driver	1.0000	1.0000	1.0000	1.0000	1.0000
	Database	0.0000	0.0000	0.0000	0.0000	1.0000
	Network	0.8286	0.7838	0.8286	0.8056	0.8841
	Human Factors	0.5000	1.0000	0.5000	0.6667	1.0000
	Design and Manufacturing	0.4286	1.0000	0.4286	0.6000	1.0000
	Performance and Calibration	0.7500	1.0000	0.7500	0.8571	1.0000
10	Software	0.7442	0.9697	0.7442	0.8421	0.9836
	Driver	1.0000	1.0000	1.0000	1.0000	1.0000
	Database	0.5000	0.5000	0.5000	0.5000	0.9825
	Network	0.8889	0.9010	0.8889	0.8950	0.9139
	Human Factors	0.5000	0.5000	0.5000	0.5000	0.9299
	Design and Manufacturing	0.6667	1.0000	0.6667	0.8000	1.0000
	Performance and Calibration	0.7778	0.7500	0.7778	0.7633	0.9786

Table 9
Results of 10-fold validation on 1048 records.

Fold	Total accuracy	Average accuracy	Precision	Recall	F1 Score
1	0.8095	0.5338	0.6167	0.5338	0.5645
2	0.7810	0.5380	0.5906	0.5380	0.5260
3	0.7714	0.5487	0.6285	0.5487	0.5733
4	0.7429	0.4671	0.4731	0.4671	0.4685
5	0.8381	0.6640	0.7169	0.6640	0.6720
6	0.7714	0.5455	0.6001	0.5455	0.5336
7	0.7714	0.4506	0.4617	0.4506	0.4523
8	0.7885	0.6469	0.6798	0.6469	0.6540
9	0.7788	0.6273	0.7858	0.6273	0.6744
10	0.7788	0.7254	0.6832	0.7155	0.6874

enabling flexibility during development. Future work may focus on automated or semi-automated ontology mapping techniques to align taxonomy classes with established clinical or regulatory terminologies (e.g., SNOMED-CT [48] or MedDRA [49]), thereby improving semantic consistency, interoperability, and integration into broader health information and regulatory ecosystems

5.2.4. Imbalance-aware learning considerations

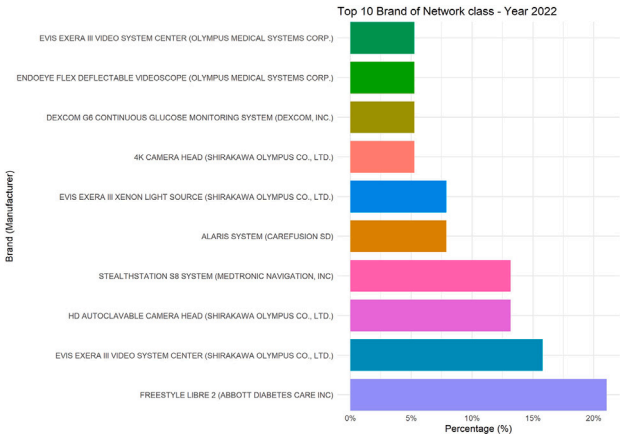
An additional limitation concerns the sample set used for fine-tuning the BERT classifier. Due to time and resource constraints, only 2000 records could be manually processed and labelled, of which just 1048 were ultimately suitable for analysis. This reduced sample size limited the availability of records for the required fine-tuning. Moreover, a pronounced class imbalance emerged within the dataset: among the nine identified classes, some were well represented, whereas others, such as *Cybersecurity* and *Service-Induced Failure*, were nearly absent or contained fewer than five examples. Although clustering-based sampling helped mitigate dominance effects during dataset construction, it does not fully compensate for severe class imbalance during model training, which hindered the model's ability to learn effectively, particularly for underrepresented classes, resulting in less reliable predictions.

To address this limitation and enhance model performance in future works, two potential strategies can be adopted. The former involves substantially increasing the quantity of labelled data, either by allocating more time to manual labelling or by expanding the workforce. A larger dataset, even with persistent class imbalance, would ensure that underrepresented classes have enough examples to support more reliable training. The latter implies using more advanced imbalance-aware strategies, such as oversampling in embedding space (e.g., SMOTE applied to contextual text embeddings), cost-sensitive learning through class-weighted loss functions, or leveraging AI-based text generation models to create synthetic records for each class. These approaches could help balance the dataset and improve representativeness. However, it requires careful consideration, as synthetic data could introduce biases or affect the validity of the resulting model.

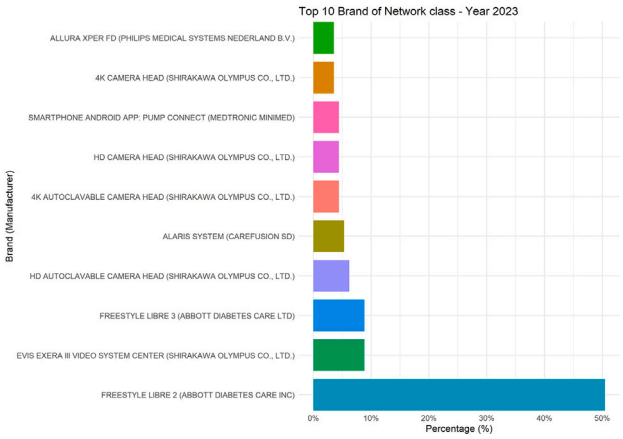
5.2.5. Explainability and regulatory transparency

Model explainability is essential to support transparency, traceability, and human oversight for possible future deployment in regulatory and clinical contexts. Although explainability techniques were not explicitly implemented in the present study, the adopted transformer-based architecture naturally supports post hoc interpretability analyses.

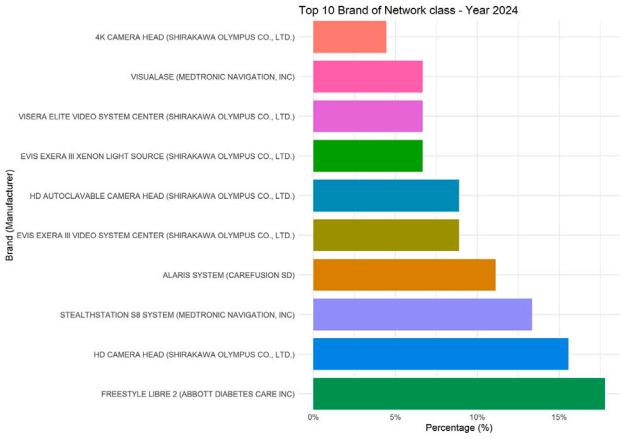
Attention-weight visualisation can be used to highlight salient tokens and textual segments that most influence classification decisions, enabling domain experts to verify whether the model focuses on clinically and technically meaningful evidence within adverse event narratives. In addition, model-agnostic explainability approaches, such as SHAP applied to contextual text embeddings, could quantify the contribution of individual words or phrases to predicted failure classes. Integrating such explainability mechanisms would enhance trust in automated failure classification, facilitate expert validation of model outputs, and support alignment with regulatory expectations for transparency and risk control in SaMD lifecycle management.



(a) Top 10 devices by frequency in the *Network* class for 2022.



(b) Top 10 devices by frequency in the *Network* class for 2023.



(c) Top 10 devices by frequency in the *Network* class for 2024.

Fig. 10. Comparison between the Top 10 devices by frequency in the *Network* class for years 2022 to 2024.

6. Conclusion

This study introduces a deep learning and NLP-based framework for classifying HIT-related software failures, specifically meant to inform risk-aware design and development of SaMD. By synthesising insights from literature-derived taxonomies, self-supervised token extraction, and expert-validated clustering, the proposed classification captures nine critical failure modes that recur across RWD extracted from the

Table 10
Mean ± standard deviation of class-specific performance metrics across 10-fold cross-validation.

Class	Accuracy	Precision	Recall	F1-score
Software	0.8523 ± 0.0472	0.8248 ± 0.0833	0.8523 ± 0.0472	0.8347 ± 0.0374
Driver	0.2500 ± 0.4249	0.3000 ± 0.4830	0.2500 ± 0.4249	0.2667 ± 0.4389
Database	0.3833 ± 0.3932	0.4250 ± 0.4091	0.3833 ± 0.3932	0.3924 ± 0.3800
Network	0.8917 ± 0.0614	0.8527 ± 0.0628	0.8917 ± 0.0614	0.8701 ± 0.0464
Human Factors	0.4036 ± 0.2393	0.5500 ± 0.3036	0.4036 ± 0.2393	0.4415 ± 0.2339
Design and Manufacturing	0.5476 ± 0.1437	0.7689 ± 0.2695	0.5476 ± 0.1437	0.6131 ± 0.1611
Performance and Calibration	0.6945 ± 0.1064	0.7780 ± 0.1778	0.6945 ± 0.1064	0.7202 ± 0.0966

Table 11
Mapping between the proposed failure taxonomy and risk management phases defined by ISO 14971 and IMDRF.

Risk management phase	Relevant taxonomy elements	Practical support provided
Hazard identification	Failure categories (e.g., Software, Network, Human Factors)	Systematic identification of recurrent failure sources from post-market surveillance data, enabling early detection of software and system-level hazards.
Risk analysis and prioritisation	Class frequency and misclassification patterns	Estimation of relative prominence and interaction of failure modes, supporting risk ranking and focus on dominant or emerging hazards.
Verification	Technology-specific failure categories (e.g., Database, Network, Performance)	Guidance for defining targeted verification activities and test cases addressing known failure mechanisms observed in real-world deployments.
Validation	Human Factors and system integration failures	Support for validation scenarios reflecting realistic use conditions, socio-technical interactions, and operational contexts.
Post-market monitoring	Longitudinal application of the taxonomy	Consistent categorisation of new adverse events, enabling trend analysis and feedback into continuous risk management processes.

MAUDE database (2022–2024). Focusing on recent MAUDE data aligns with the study's objective of supporting contemporary, risk-informed SaMD design, where insights derived from current software architectures, regulatory expectations, and clinical deployment contexts are most actionable for manufacturers. Moreover, in this initial study, conservative data handling strategies were preferred over aggressive rebalancing to avoid introducing artefacts into safety-critical failure categories.

The multinomial BERT classifier achieved robust performance in cross-validation (total accuracy 74.29–83.81%, F1-score up to 0.87 for dominant classes), demonstrating the feasibility of integrating advanced NLP pipelines into regulatory-aligned SaMD development workflows. Nonetheless, several limitations remain. The development of a BERT-based model for automated classification of report texts faced substantial challenges due to class imbalance and the limited availability of representative samples. While preliminary outcomes are promising, this aspect requires further investigation and refinement. Specifically, the manual labelling process underscored the need to improve the quality and detail of failure reporting. Many records lacked sufficient information and rarely included troubleshooting details, which is especially problematic for software-related issues, where brief descriptions are inadequate for accurate assessment. The predominance of generic categories, such as *Network* or *Software*, highlights the importance of enhancing the technical training of maintenance personnel, particularly in IT (Information Technology) and ICT (Information and Communication Technology) domains, to ensure that future failure reports are more detailed and actionable.

The proposed framework supports manufacturers in:

- anticipating software risks during design and development phases
- informing hazard analyses and verification strategies consistent with ISO 14971 and IEC 62304 standards [5,6]
- enhancing post-market surveillance by enabling automated monitoring and trend detection
- establishing a continuous feedback loop where RWE informs iterative design improvements.

Moreover, it introduces a data-driven approach in the lifecycle management of SaMD based on RWE, bridging pre-market design, risk control, and post-market vigilance to foster safer, more reliable biomedical software technologies.

CRediT authorship contribution statement

Alessio Luschi: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Data curation, Conceptualization. **Alessandro Zazzeri:** Validation, Software, Investigation, Formal analysis, Data curation. **Gabriele Cevenini:** Writing – review & editing, Visualization, Conceptualization. **Ernesto Iadanza:** Writing – review & editing, Supervision, Project administration, Conceptualization.

Funding

This research received no external funding.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] R. Sinha, The role and impact of new technologies on healthcare systems, Discov. Health Syst. 3 (2024).
[2] A. Benhanifia, Z.B. Cheikh, P.M. Oliveira, A. Valente, J. Lima, Systematic review of predictive maintenance practices in the manufacturing sector, Intell. Syst. Appl. 26 (2025) 200501.
[3] European Union, Medical devices regulation (EU) 2017/745, 2017, https://eur-lex.europa.eu/legal-content/IT/ALL/?uri=celex:32017R0745, (Accessed 23 July 2025).
[4] International Medical Device Regulators Forum, IMDRF terminologies for categorized adverse event reporting (AER): Terms, terminology structure and codes, 2020, https://www.imdrf.org/documents/terminologies-categorized-adverse-event-reporting-aer-terms-terminology-and-codes, (Accessed 23 March 2023).
[5] International Organization for Standardization, ISO 14971:2019, Medical devices — Application of risk management to medical devices, 2019.

[6] International Electrotechnical Commission, IEC 62304:2006, Medical device software — Software life cycle processes, 2006.
[7] L. Mascii, A. Luschi, E. Iadanza, Sentiment analysis for performance evaluation of maintenance in healthcare, in: IFMBE Proceedings, Vol. 84, 2021, pp. 359–367.
[8] E. Iadanza, S. Cerofolini, C. Lombardo, F. Satta, M. Gherardelli, Medical devices nomenclature systems: a scoping review, Health Technol. 11 (2021) 681–692.
[9] World Health Organization, Nomenclature of medical devices, 2023, https://www.who.int/teams/health-product-policy-and-standards/assistive-and-medical-technology/medical-devices/nomenclature, (Accessed 23 July 2025).
[10] E. Iadanza, A. Luschi, Standardization of failure codes and nomenclature of medical devices for evidence-based maintenance, in: IFMBE Proceedings, Vol. 94, 2024, pp. 170–177.
[11] A. Luschi, C. Petraccone, G. Fico, L. Pecchia, E. Iadanza, Semantic ontologies for complex healthcare structures: A scoping review, IEEE Access 11 (2023) 19228–19246.
[12] E. Iadanza, V. Gonnelli, F. Satta, M. Gherardelli, Evidence-based medical equipment management: a convenient implementation, Med. Biol. Eng. Comput. 57 (2019) 2215–2230.
[13] E. Iadanza, L. Marzi, F. Dori, G. Biffi Gentili, M. Torricelli, Hospital health care offer. A monitoring multidisciplinar approach, in: IFMBE Proceedings, Vol. 14, 2007, pp. 3685–3688.
[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017.
[15] J. Devlin, M.-W. Chang, K. Lee, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.
[16] M.-Y. You, F. Liu, W. Wang, G. Meng, Statistically planned and individually improved predictive maintenance management for continuously monitored degrading systems, IEEE Trans. Reliab. 59 (4) (2010) 744–753.
[17] H. Kang, J. Wang, B. Yao, S. Zhou, Y. Gong, Toward safer health care: a review strategy of FDA medical device adverse event database to identify and categorize health information technology related events, JAMIA Open 2 (1) (2018) 179–186.
[18] F. Crapanzano, A. Luschi, F. Satta, L. Sani, E. Iadanza, Evidence based management of medical devices: A follow-up experiment, Biomed. Signal Process. Control 99 (2025) 106867.
[19] A. Luschi, P. Nesi, E. Iadanza, Evidence-based clinical engineering: Health information technology adverse events identification and classification with natural language processing, Heliyon 9 (11) (2023) e21723.
[20] E. Denisova, E. Tiribilli, A. Luschi, P. Francia, L. Manetti, L. Bocchi, E. Iadanza, Enabling reliable usability assessment and comparative analysis of medical software: a comprehensive framework for multimodal biomedical imaging platforms, Health Technol. 14 (4) (2024) 671–682.
[21] A.L.D. Neves, L.E.G. Martins, M.A.L. Gabbay, G. Cavicchioli, F.S. Tenorio, T.S. Cunha, Insulin pump-associated adverse events in a Brazilian reference center for the treatment of diabetes mellitus: Proposal for a taxonomy of device failures in adults, adolescents, and children, J. Diabetes Sci. Technol. 18 (1) (2024) 74–81.
[22] S.W. Hassanin, R.S. Kshirsagar, J.G. Eide, J. Chang, J. Liang, J.N. Palmer, N.D. Adappa, Image-guided surgical device failures in functional endoscopic sinus surgery: A MAUDE analysis, Laryngoscope 133 (6) (2023) 1310–1314.
[23] S.H. Tak, H. Choi, D. Lee, Y.A. Song, J. Park, Nurses' perceptions about smart beds in hospitals, CIN: Comput. Inform. Nurs. 41 (6) (2023).
[24] F. Castagnini, M. Maestri, E. Tassinari, C. Masetti, C. Faldini, F. Traina, Software-related recalls in computer-assisted hip and knee arthroplasty, Int. Orthop. 47 (3) (2023) 641–645.
[25] X. Qian, C.J. Channels, S.A. Gaeta, M.H. Wish, B. Matthews, B.D. Atwater, V. Kumar, Radiofrequency remote monitor software patch update without cybersecurity implantable cardioverter-defibrillator firmware update increases the risk of inappropriate implantable cardioverter-defibrillator therapies, HeartRhythm Case Rep. 8 (2) (2022) 69–72.
[26] A. Radovici, I. Culic, D. Rosner, F. Oprea, A model for the remote deployment, update, and safe recovery for commercial sensor-based IoT systems, Sensors 20 (16) (2020) 4393.
[27] J.Z. Lee, M.J. Henrich, P. Bibby, S.K. Mulpuru, P.A. Friedman, Y.-M. Cha, K. Srivathsan, Pacemaker firmware update and interrogation malfunction, HeartRhythm Case Rep. 5 (4) (2019) 213–216.
[28] B. Cowie, L. Baker, B. Shoghi, M. Worner, D. Scott, Electrocardiogram failure in the operating room – bench testing to prevent bed-side disaster, Anaesthesia 73 (6) (2018) 746–749.
[29] G. Chu, V. Li, A. Hui, C. Lam, E. Chan, M. Law, L. Yip, W. Lam, Failure analysis for ultrasound machines in a radiology department after implementation of predictive maintenance method, J. Med. Ultrasound 26 (1) (2018).
[30] M.J. Connor, D.C. Marshall, V. Moiseenko, K. Moore, L. Cervino, T. Atwood, P. Sanghvi, A.J. Mundt, T. Pawlicki, A. Recht, J.A. Hattangadi-Gluth, Adverse events involving radiation oncology medical devices: Comprehensive analysis of US food and drug administration data, 1991 to 2015, Int. J. Radiat. Oncol. Biol. Phys. 97 (1) (2017) 18–26.
[31] E. Rajih, C. Tholomier, B. Cormier, V. Samouëlian, T. Warkus, M. Liberman, H. Widmer, J.-B. Lattouf, A.M. Alenizi, M. Meskawi, et al., Error reporting from the da Vinci surgical system in robotic surgery: A Canadian multispecialty experience at a single academic centre, Can. Urol. Assoc. J. 11 (5) (2017) E197–202.
[32] J.G. Ronquillo, D.M. Zuckerman, Software-related recalls of health information technology and other medical devices: Implications for FDA regulation of digital health, Milbank Q. 95 (3) (2017) 535–553.
[33] A. Wright, T.-T.T. Hickman, D. McEvoy, S. Aaron, A. Ai, J.M. Andersen, S. Hussain, R. Ramoni, J. Fiskio, D.F. Sittig, D.W. Bates, Analysis of clinical decision support system malfunctions: a case series and survey, J. Am. Med. Inform. Assoc. 23 (6) (2016) 1068–1076.
[34] A. Ferrarese, G. Pozzi, F. Borghi, A. Marano, P. Delbon, B. Amato, M. Santangelo, C. Buccelli, M. Niola, V. Martino, E. Capasso, Malfunctions of robotic system in surgery: role and responsibility of surgeon in legal point of view, Open Med. (Warsaw Pol.) 11 (1) (2016) 286–291.
[35] F. Pecoraro, D. Luzi, Detecting software failures in the MAUDE database: a preliminary analysis, Stud. Health Technol. Inform. 192 (2013) 1098.
[36] G. Chung, K. Etter, A. Yoo, Medical device active surveillance of spontaneous reports: A literature review of signal detection methods, Pharmacoepidemiol. Drug Saf. 29 (4) (2020) 369–379.
[37] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS '17, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 4768–4777.
[38] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, 2016, pp. 1135–1144.
[39] Q. Zhou, B. Sun, Adaptive K-means clustering based under-sampling methods to solve the class imbalance problem, Data Inf. Manag. 8 (3) (2024) 100064.
[40] M. Goyal, A review on K-mode clustering algorithm, Int. J. Adv. Res. Comput. Sci. 8 (2017) 725–729.
[41] R. Blagus, L. Lusa, SMOTE for high-dimensional class-imbalanced data, BMC Bioinformatics 14 (1) (2013) 106.
[42] J.M. Johnson, T.M. Khoshgoftaar, Survey on deep learning with class imbalance, J. Big Data 6 (1) (2019) 27.
[43] C.J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, D. King, Key challenges for delivering clinical impact with artificial intelligence, BMC Med. 17 (1) (2019) 195.
[44] A. Bookstein, V.A. Kulyukin, T. Raita, Generalized hamming distance, Inf. Retr. 5 (4) (2002) 353–375.
[45] M. Schreier, Qualitative Content Analysis in Practice, SAGE Publications Ltd, 2012.
[46] M.L. McHugh, Interrater reliability: the kappa statistic, Biochem. Medica 22 (3) (2012) 276–282.
[47] S. El Anigri, M.M. Himmi, A. Mahmoudi, How BERT's dropout fine-tuning affects text classification? in: M. Fakir, M. Baslam, R. El Ayachi (Eds.), Business Intelligence, Springer International Publishing, Cham, 2021, pp. 130–139.
[48] SNOMED International, Systematized nomenclature of medicine clinical terms, 2021, https://www.snomed.org/, (Accessed 13 January 2026).
[49] International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, Medical dictionary for regulatory activities, 2025, https://www.meddra.org/, (Accessed 13 January 2026).