

Uncertainty-Aware Transformer-Centred Medical Text Classification: A Systematic Methodological Review and Taxonomy

Prabhashrini Dhanushika Manage

`prabhashrinidhanushika.manage@hdr.qut.edu.au`

Queensland University of Technology

Yutong Wu

The Australian e-Health Research Centre, CSIRO

Jinglan Zhang

Queensland University of Technology

Yuefeng Li

Queensland University of Technology

Research Article

Keywords: Transformer, BERT, Medical text classification, Uncertainty-aware decision making, Biomedical NLP, Systematic methodological review

Posted Date: October 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10970629/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Uncertainty-Aware Transformer-Centred Medical Text Classification: A Systematic Methodological Review and Taxonomy

Prabhashrini Dhanushika Manage^{1,2}, Yutong Wu³,
Jinglan Zhang¹, Yuefeng Li^{1*}

¹School of Computer Science, Queensland University of Technology,
Brisbane, 4000, QLD, Australia .

²Department of Computational Mathematics, University of Moratuwa,
Sri Lanka .

³Australian e-Health Research Centre, CSIRO, Brisbane, 4029, QLD,
Australia .

*Corresponding author. E-mail: y2.li@qut.edu.au;

Contributing authors: prabhashrinidhanushika.manage@hdr.qut.edu.au;
dhanushikam@uom.lk; yutong.wu@csiro.au; jinglan.zhang@qut.edu.au;

Abstract

Transformer-based medical text classifiers can achieve strong predictive performance yet remain overconfident for ambiguous, noisy, or weakly supported instances, limiting their reliability in safety-critical settings. This systematic methodological review addresses the lack of a dedicated functional synthesis of uncertainty-aware methods in transformer-centred medical text classification. Following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020, six literature sources were searched in May 2026, identifying 5,895 records; 32 primary studies met the eligibility criteria. We introduce a four-tier functional taxonomy that organises reliability interventions by where they act in the classification pipeline: feature representation, prediction-level uncertainty estimation, uncertainty-aware decision making, and training-data reliability. The 32 studies yielded 45 taxonomy assignments because 13 studies contributed substantively to two tiers. Feature representation was the largest category (21/45, 46.7%), followed by training-data reliability (12/45, 26.7%), uncertainty-aware decision making (7/45, 15.6%), and uncertainty estimation (5/45, 11.1%). Across the literature, representation and training-data interventions were substantially

more common than explicit uncertainty estimation or uncertainty-informed decision policies. Evaluation was also fragmented: uncertainty measures, decision rules, and training-data reliability criteria varied across datasets and settings, and no shared validation framework was identified. The proposed taxonomy provides a unified methodological basis for comparing otherwise heterogeneous approaches and highlights three priorities for reliable transformer-based medical text classification: consistent prediction-level uncertainty evaluation, cross-setting validation of decision rules, and transparent validation of training evidence.

Keywords: Transformer, BERT, Medical text classification, Uncertainty-aware decision making, Biomedical NLP, Systematic methodological review

1 Introduction

The automated classification of medical text supports applications ranging from diagnostic coding and clinical decision support to pharmacovigilance and biomedical literature synthesis. Reliable classification is particularly important in healthcare, where misclassifications may affect clinical decision-making and trust in model outputs (Loftus et al. 2022). Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2019) substantially advanced text classification, while domain-adapted variants such as BioBERT (Lee et al. 2020), ClinicalBERT (Huang et al. 2019), PubMedBERT (Gu et al. 2022), SciBERT (Beltagy et al. 2019), and BioALBERT (Naseem et al. 2022) improved representations for biomedical and clinical language through domain-specific pre-training.

Despite these advances, transformer-based classifiers can remain overconfident for ambiguous, noisy, or weakly supported instances. Standard fine-tuning with a softmax classification head may assign high confidence even to misclassified or distributionally atypical examples (Hendrycks and Gimpel 2017), while uncertainty estimates from pre-trained language models remain imperfect in safety-critical settings (Wang et al. 2022). In medical applications, uncertainty may affect diagnosis, treatment planning, and clinical decision-making (Alizadehsani et al. 2024), while ambiguity in clinical terminology further complicates reliable interpretation of medical text (Link et al. 2022). These difficulties arise from both aleatoric uncertainty, including genuine semantic ambiguity, and epistemic uncertainty associated with limited model knowledge or training evidence (Alizadehsani et al. 2024). Pre-trained transformers may also remain overconfident under noisy text classification conditions (Zhang et al. 2023a). Consequently, high predicted confidence does not necessarily indicate reliable classification for difficult medical-text instances.

Existing surveys have examined uncertainty estimation in healthcare, uncertainty in natural language processing more broadly, transformer architectures for medical natural language processing (NLP), and individual uncertainty-aware techniques. However, to our knowledge, no systematic methodological review has provided a dedicated functional synthesis of uncertainty-aware methods in transformer-centred

medical text classification. The literature spans several distinct forms of reliability intervention, including strengthening the evidence represented for classification, estimating prediction-level uncertainty, using reliability signals to determine inference-time actions, and controlling the quality of training evidence. Considering these mechanisms within a common framework is important because improved classification performance alone does not establish calibrated or operationally reliable predictions.

This review therefore aims to identify, classify, and critically synthesise uncertainty-aware methods in transformer-centred medical text classification, with particular attention to where reliability-related mechanisms intervene, how uncertainty and reliability are evaluated, and what methodological gaps remain. The review is transformer-centred rather than transformer-exclusive: BERT and its domain-adapted variants form the principal methodological focus, while three principally non-transformer studies were retained because they address directly comparable reliability functions within medical text classification. Accordingly, the term uncertainty-aware is used here in a functional, reliability-oriented sense, encompassing not only explicit prediction-level uncertainty estimation but also mechanisms that strengthen the evidence used for classification, use reliability signals to guide decisions, or control the quality of training evidence.

The review makes four main contributions:

- We conduct a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)-guided systematic methodological review of 32 primary studies published between 2019 and May 2026, examining reliability-related methods in medical, clinical, biomedical, and health-related text classification.
- We propose a four-tier functional taxonomy that organises uncertainty-aware methods according to *where* their reliability-related mechanism intervenes in the medical text-classification pipeline. Tier 1, *feature representation*, covers methods that strengthen or enrich the evidence provided to the classifier, including probabilistic class-conditional features, biomedical knowledge, domain-specific pre-training, feature fusion, attention, and long-context modelling. Tier 2, *uncertainty estimation*, covers methods that explicitly quantify or evaluate prediction-level uncertainty, confidence, calibration, or model-output reliability. Tier 3, *uncertainty-aware decision making*, covers methods that use uncertainty, confidence, prediction stability, or model agreement to determine inference-time actions such as acceptance, routing, fallback, early exiting, rejection, deferral, or abstention. Tier 4, *training-data reliability*, covers methods that control the construction, selection, labelling, filtering, weighting, or influence of training evidence. Studies making substantive contributions at two distinct intervention points are assigned to both applicable tiers.
- We compare the included studies according to medical domain, dataset, classification backbone, reliability-related mechanism, evaluation approach, taxonomy assignment, and principal reported finding. The 32 unique studies yield 45 taxonomy assignments.
- We identify methodological gaps concerning the limited evaluation of representation improvements using prediction-level reliability measures, limited

cross-setting validation of uncertainty-aware decision rules, and the absence of a shared validation framework for training-data reliability.

In addition to the 32 primary studies included in the systematic synthesis, methodologically relevant secondary sources were retained only to provide theoretical, medical, and comparative context and were excluded from the primary-study counts, taxonomy distribution, and comparative analysis.

2 Related Work

2.1 BERT and Domain-Adapted Variants for Medical NLP

BERT (Devlin et al. 2019) provides bidirectional contextual representations widely used for text classification. Because medical language contains specialised terminology, abbreviations, and semantic patterns, domain-specific pre-training has produced variants including BioBERT (Lee et al. 2020), ClinicalBERT (Huang et al. 2019), PubMedBERT (Gu et al. 2022), SciBERT (Beltagy et al. 2019), and BioALBERT (Naseem et al. 2022).

Domain adaptation does not eliminate reliability concerns. BERT performance can vary across domains and tasks (Coban et al. 2024), and clinical classification remains dataset- and task-dependent (Gao et al. 2021). Attention may also emphasise contextual patterns that are not uniformly informative (Sun and Lu 2020; Reusens et al. 2024), while fragmentation of specialised terminology motivates biomedical pre-training such as BioBERT and PubMedBERT (Lee et al. 2020; Gu et al. 2022). Moreover, softmax classifiers may remain poorly calibrated or overconfident for ambiguous, noisy, or boundary instances (Guo et al. 2017; Desai and Durrett 2020; Zhang et al. 2023a). Improved domain representation therefore does not by itself establish reliable prediction confidence.

2.2 Types of Uncertainty and Foundational Methods

Epistemic uncertainty reflects limitations in model knowledge, training evidence, or capacity, whereas aleatoric uncertainty arises from properties of the data, including semantic ambiguity, overlapping classes, inconsistent documentation, or annotation noise (Alizadehsani et al. 2024). These sources may interact in medical text.

Softmax confidence is commonly used as a proxy for prediction reliability and for identifying misclassified or out-of-distribution examples (Hendrycks and Gimpel 2017), although pretrained transformers may remain poorly calibrated under noisy text or distribution shift (Zhang et al. 2023a). Relevant foundational approaches include Monte Carlo dropout (Gal and Ghahramani 2016; Islam et al. 2022), evidential deep learning (Sensoy et al. 2018; Hu and Khan 2021), deep ensembles (Lakshminarayanan et al. 2017), calibration methods (Guo et al. 2017; Desai and Durrett 2020), conformal prediction (Campos et al. 2024), and three-way decision theory (Li et al. 2017). These methods provide context for the reliability mechanisms examined in the primary synthesis.

2.3 Methodologically Relevant Studies Retained as Background

Several relevant studies were retained as background but excluded from primary-study counts, taxonomy assignments, and comparative analysis. [Dong et al. \(2025\)](#) reviewed calibration for class-imbalanced deep-learning classifiers; [Liu et al. \(2025\)](#) surveyed uncertainty quantification and confidence calibration in large language models; and [Ulmer et al. \(2023\)](#) reviewed evidential uncertainty-estimation approaches. These were retained as secondary methodological context rather than eligible primary studies.

Other studies clarify eligibility boundaries. [Moradi and Samwald \(2021\)](#) addressed post-hoc explanation without predictive-uncertainty estimation; [Solarte-Pabón et al. \(2022\)](#) examined token-level negation and linguistic-uncertainty cues rather than classifier uncertainty; and [Ferrell et al. \(2023\)](#) studied research-protocol classification outside the eligible medical-text scope. [Tajrian et al. \(2026\)](#) reported bootstrapped confidence intervals for aggregate classifier performance rather than prediction-level uncertainty. [Zhang et al. \(2026\)](#) combined structured primary-care variables with TF-IDF clinical-note features but did not estimate predictive uncertainty, evaluate calibration, or use uncertainty to control classification decisions. These studies therefore remain secondary context rather than part of the primary four-tier synthesis.

3 Methods

3.1 Search Strategy

This systematic review followed the PRISMA 2020 guidelines ([Page et al. 2021](#)). Searches were conducted in May 2026 across six literature sources: PubMed, IEEE Xplore, ACM Digital Library, Scopus, Web of Science, and arXiv via Semantic Scholar, covering publications from January 2013 to May 2026. Database-specific search strategies combined four conceptual blocks: (i) transformer-based language models; (ii) text classification; (iii) uncertainty and confidence; and (iv) the medical domain. Table 1 reports the complete search strings, searched fields, and numbers of records identified.

The searches identified 5,895 records, of which 4,488 were successfully imported into Mendeley Reference Manager because of platform export limitations. ACM Digital Library exported 934 of 1,335 records, Scopus 3,466 of 3,570, and IEEE Xplore 32 of 37; the five unimported IEEE records were book chapters. The 897 arXiv records identified through Semantic Scholar could not be bulk-exported separately; relevant preprints were considered when retrieved through another indexed source or when a peer-reviewed version was identified.

Table 1 Database-specific search strings, fields searched, and records identified (May 2026).

Database	Search String	Field(s)	Records Identified
PubMed	(BERT OR BioBERT OR transformer) AND (“text classification” OR “medical text”) AND (uncertainty OR confidence OR “decision boundary”)	Title/Abstract	11
IEEE Xplore	(BERT OR BioBERT OR transformer) AND (“text classification” OR “medical text”) AND (uncertainty OR confidence OR “decision boundary”) AND (medical OR clinical OR biomedical OR health)	All Metadata	37
ACM Digital Library	(BERT OR BioBERT OR transformer OR “language model”) AND (“text classification” OR “medical text”) AND (uncertainty OR “confidence score” OR “decision boundary”)	All Fields	1,335
Scopus	(BERT OR BioBERT OR transformer) AND (“text classification” OR “medical text”) AND (uncertainty OR confidence OR “decision boundary”) AND (medical OR clinical OR biomedical OR health) AND PUBYEAR > 2012 AND PUBYEAR < 2027	Title, Abstract, Keywords	3,570
Web of Science	(BERT OR transformer OR BioBERT) AND (“text classification”) AND (uncertainty OR confidence OR “decision boundary”)	Abstract (TS=)	45
arXiv (Semantic Scholar)	BERT uncertainty medical text classification	All fields	897
Total			5,895

3.2 Inclusion and Exclusion Criteria

Eligibility was assessed using the predefined criteria in Table 2. Eligible studies addressed sample-level classification of medical, clinical, biomedical, or health-related text, reported quantitative classification results, and made a substantive methodological contribution to at least one of the four reliability functions represented in the taxonomy.

Table 2: Eligibility criteria for the systematic review.

Code	Criterion
Inclusion criteria	
IC1	The study evaluated classification of medical, clinical, biomedical, or health-related text as its primary task. Eligible textual inputs included clinical notes, electronic medical records, pathology or radiology reports, biomedical publications, clinical-trial protocols, health webpages, patient-generated text, and other health-related documents.

Code	Criterion
IC2	The study used a transformer-based language model as part of the proposed classification framework, or presented a directly comparable non-transformer method addressing the same reliability function within medical text classification.
IC3	The study made a substantive methodological contribution to at least one of the following functional categories: (a) feature representation through a non-standard mechanism intended to strengthen medically or class-relevant evidence; (b) prediction-level uncertainty estimation, confidence assessment, or calibration; (c) uncertainty-aware decision making, including acceptance, routing, fallback, deferral, reassessment, or abstention; or (d) training-data reliability, including synthetic-data filtering, weak-label refinement, automated labelling, active-learning selection, targeted annotation, or analysis of label quality.
IC4	The study reported quantitative classification results on at least one clearly described evaluation dataset.
IC5	The study was available as a full peer-reviewed journal article or conference paper by May 2026.

Exclusion criteria

FE1	The study focused exclusively on image, signal, or structured-data classification without a substantive text-classification component.
FE2	The primary evaluation task used exclusively non-medical data or concerned administrative classification of scientific or research documents rather than medical, clinical, biomedical, or health-related text.
FE3	The study applied a standard classifier or representation without a substantive contribution to feature reliability, predictive uncertainty estimation, uncertainty-aware decision making, or training-data reliability.
FE4	The study focused solely on named entity recognition, relation extraction, question answering, text generation, or token-level negation or linguistic-uncertainty cue detection, without a qualifying sample-level classification or training-data-reliability contribution.
FE5	The study addressed only post-hoc explainability without estimating predictive uncertainty or introducing an uncertainty-related classification or training-data mechanism.
FE6	Uncertainty was represented only through confidence intervals, bootstrap intervals, or significance tests around aggregate performance measures, without prediction-level uncertainty, calibration, uncertainty-aware decision making, or training-data reliability.
FE7	The study mentioned uncertainty, confidence, ambiguity, or reliability only superficially and did not provide a corresponding methodological or quantitative evaluation component.

Code	Criterion
FE8	The study did not report quantitative classification results or did not provide sufficient methodological information to assess eligibility.
FE9	The publication was a survey, review, editorial, commentary, or other secondary article. Relevant review papers were retained only as background references.
FE10	The record duplicated another included publication reporting the same study and results.
FE11	The work was available only as a non-peer-reviewed preprint by May 2026.

Eligibility decisions were based on each study’s primary dataset, classification task, proposed methodological mechanism, and reported evaluation. Use of a transformer alone was insufficient for inclusion, and the term “uncertainty” qualified only when the study explicitly estimated, evaluated, or operationalised model-output uncertainty, or used uncertainty-related information to improve training-data reliability.

3.3 Study Selection and Data Extraction

Title and abstract screening, full-text eligibility assessment, data extraction, and initial taxonomy coding were conducted principally by the first author and were not independently duplicated by a second reviewer. The review and synthesis were undertaken under the supervision of the second and last authors, with borderline eligibility or taxonomy-classification decisions discussed where necessary. Eligibility was assessed using Table 2. Mendeley Reference Manager was used for reference management and duplicate identification; no automated tool was used to make inclusion or exclusion decisions.

Eligible studies were examined individually using a structured data-extraction table, and extracted information was checked against the corresponding primary study report. Variables included publication year, medical or health domain, dataset, textual input, classification backbone, reliability-related mechanism, evaluation measures, taxonomy assignment, and principal reported finding. Study authors were not contacted for additional or missing information.

Extracted outcomes included conventional classification measures; prediction-reliability measures such as entropy, expected calibration error, Brier score, negative log-likelihood, and risk-coverage behaviour; operational outcomes including coverage, rejection, abstention, early exiting, automatic processing, and human review; and training-data reliability outcomes including label agreement, annotation efficiency, filtering performance, and downstream classification effects.

Three included studies did not use a transformer as the principal method proposed by the authors: [Peluso et al. \(2024\)](#), [Flores et al. \(2021\)](#), and [Flores and Verschae \(2025\)](#). They were retained because they addressed directly comparable reliability functions within medical text classification. Hybrid studies were treated as

transformer-centred when the transformer formed a substantive component of the proposed method and contributed directly to the reported primary classification results rather than serving only as an external comparator.

3.4 Taxonomy Development and Study Classification

The four-tier taxonomy was developed through study-by-study coding of the mechanisms reported in the included papers. For each study, we identified (i) the component of the classification pipeline modified by the proposed method, (ii) whether uncertainty or reliability was represented, explicitly quantified, or used to determine a subsequent action, and (iii) whether the mechanism operated during inference or training-data construction. Studies were classified according to the principal *functional intervention point* of the proposed method rather than model architecture or terminology used by the original authors.

This process identified four distinct reliability functions: strengthening the evidence available to the classifier, quantifying prediction reliability, using reliability information to determine an inference-time action, and controlling the quality or influence of training evidence. These functions define the four analytical tiers used in this review.

3.4.1 Tier 1: Feature Representation

Tier 1 includes substantive mechanisms that strengthen class- or medically relevant representations, including probabilistic class-conditional features, biomedical knowledge augmentation, domain-specific pre-training, long-context encoding, multi-task learning, attention, and structured feature fusion. Standard transformer representations alone were insufficient for Tier 1 assignment, and representation improvements were not classified as prediction-level uncertainty estimation unless output reliability, confidence, calibration, or uncertainty was explicitly quantified.

3.4.2 Tier 2: Uncertainty Estimation

Tier 2 includes methods that explicitly quantify or evaluate prediction-level uncertainty, confidence, calibration, or model-output reliability, including Bayesian approximations, predictive entropy, Brier score, negative log-likelihood, expected calibration error, and related measures. Confidence intervals or significance tests applied only to aggregate performance were not treated as prediction-level uncertainty.

3.4.3 Tier 3: Uncertainty-aware decision making

Tier 3 includes methods in which uncertainty, confidence, prediction stability, or model agreement determines an inference-time action such as acceptance, routing, fall-back, continued computation, early exiting, rejection, human review, reassessment, or abstention. Tier 3 therefore differs from Tier 2 by requiring the reliability signal to influence what the system subsequently does with a prediction.

3.4.4 Tier 4: Training-Data Reliability

Tier 4 includes mechanisms controlling the quality, selection, construction, weighting, or influence of training evidence, including report pre-selection, synthetic-data generation and filtering, noise-tolerant training, weak-label refinement, silver- and gold-label strategies, automated labelling, active-learning selection, and targeted annotation.

Studies making substantive evaluated contributions at two distinct functional intervention points were assigned to both applicable tiers; the presence of a secondary technique alone was insufficient. No study was assigned to more than two tiers. Applying these rules to the 32 included studies produced 45 assignments because 13 studies contributed substantively to two tiers. Final assignments therefore reflected each method’s evaluated functional role in the classification pipeline rather than the terminology used by the original authors.

3.5 Synthesis Methods

Studies were synthesised according to the taxonomy-assignment rules in Section 3.4. Because the included studies differed substantially in medical domains, datasets, languages, classification tasks, model architectures, and definitions and measures of uncertainty or reliability, quantitative pooling was not considered appropriate for the aims of this review. Findings were therefore synthesised narratively across the four functional tiers.

Study characteristics, methodological contributions, evaluation approaches, taxonomy assignments, and principal reported findings were summarised in the comparative analysis table. Taxonomy distribution was described using counts and percentages of tier assignments. Quantitative results were retained as reported by the original authors; missing statistics were not imputed and statistical conversions were not performed.

Statistical heterogeneity was not quantified, and no subgroup analyses, meta-regression, or sensitivity analyses were conducted. Formal study-level risk-of-bias, reporting-bias, and certainty-of-evidence assessments were not undertaken because the review focused on functional methodological synthesis rather than estimation of comparative intervention effects. The findings are therefore interpreted as a descriptive methodological synthesis rather than a formally graded body of evidence.

4 Results

4.1 Study Selection and Taxonomy Distribution

The searches identified 5,895 records, of which 4,488 were successfully imported into Mendeley Reference Manager owing to database-export limitations. After removal of 31 duplicates, 4,457 unique records were screened, and 4,363 were excluded. Three further bibliographic duplicates were removed under FE10 from the 94 remaining reports, leaving 91 for full-text eligibility assessment.

Of the 91 full-text reports, 32 unique studies met the eligibility criteria and were included in the primary synthesis. Other reports were either formally excluded under the predefined criteria or retained as secondary background sources; the latter

were excluded from primary-study counts, taxonomy assignments, and comparative analysis.

Study-level reasons for formal exclusion are reported in Appendix Table 4, and the study-selection process is shown in Figure 1.

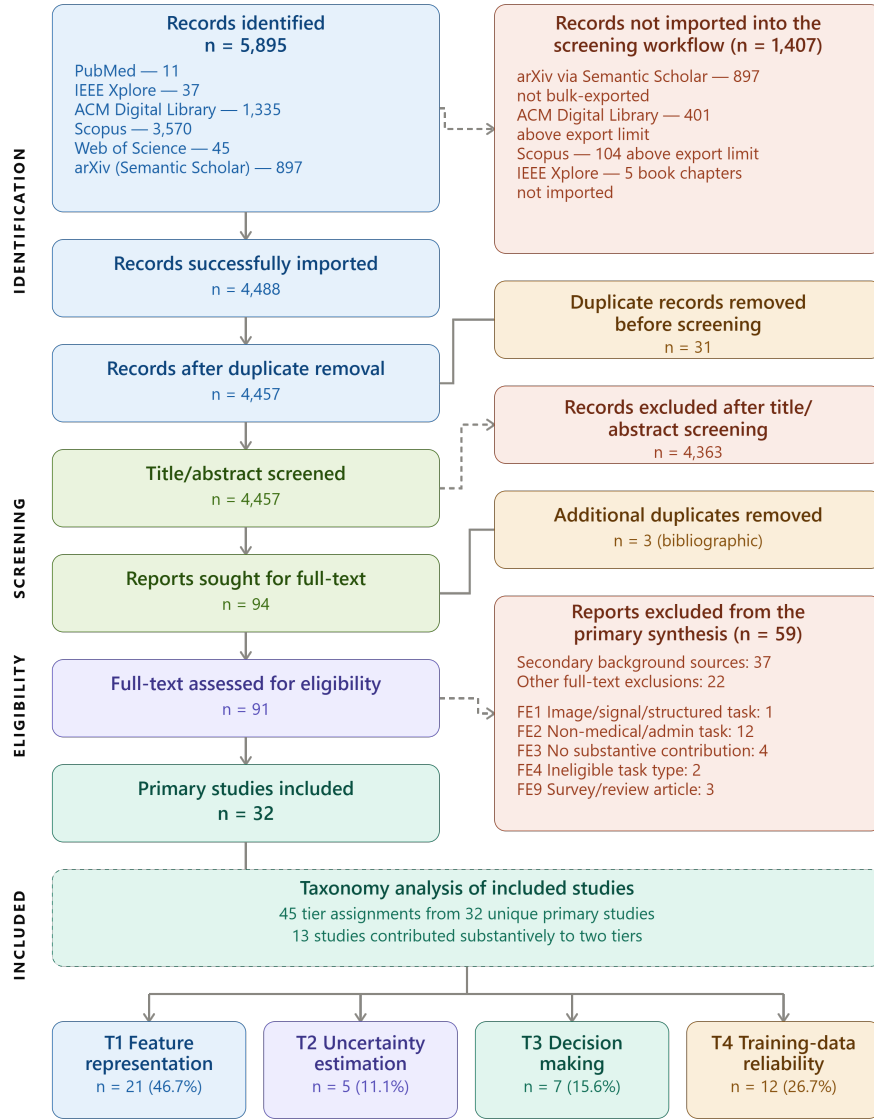


Fig. 1 PRISMA 2020 flow diagram for the identification, screening, eligibility assessment, and selection of 32 primary studies.

The 32 studies produced 45 taxonomy assignments because 13 studies contributed substantively to two tiers. Tier 1 was the largest category with 21 assignments (46.7%), followed by Tier 4 with 12 (26.7%), Tier 3 with 7 (15.6%), and Tier 2 with 5 (11.1%). No study was assigned to more than two tiers.

4.2 Tier 1: Feature Representation Methods ($n = 21$)

Tier 1 comprises methods that strengthen the evidence available to the classifier through probabilistic class-conditional information, biomedical knowledge, domain-specific pre-training, feature fusion, attention, multi-task learning, long-context modelling, or task-specific features. It accounts for 21/45 (46.7%) of all taxonomy assignments; use of a standard transformer representation alone was insufficient for assignment.

4.2.1 Probabilistic Term-Class Augmentation

[Manage et al. \(2026b\)](#) introduced Term-Class Probability Embedding (TCPE), combining smoothed term-class conditional probabilities with TF-IDF weighting and BERT’s [CLS] representation. On PubMed 20K and Ohsumed, SVD visualisations showed more compact and separated class structures.

[Manage et al. \(2026a\)](#) subsequently introduced Word-Class Probabilistic Embedding (WCPE) within a two-stage framework. Average probability difference (AvgDiff) separates confident and uncertain samples, with WCPE-enhanced representations selectively applied to uncertain instances. Together, TCPE and WCPE explicitly inject class-conditional probabilistic evidence into document representations.

4.2.2 Domain-Enriched and Knowledge-Augmented Architectures

[Saqib and Mustafa \(2025\)](#) augmented PubMedBERT with UMLS Concept Unique Identifiers from clinical notes. NER-filtered concepts improved diagnosis classification, particularly for minority classes, whereas unfiltered augmentation introduced noise.

[Rostam and Kertész \(2026\)](#) enriched PubMed abstract sentences with scispaCy biomedical entities and evaluated BERT, SciBERT, BioBERT, and attention-based multiple-instance learning for EQ-5D study identification, contributing entity-enriched sentence- and study-level representations.

[Alnajjar et al. \(2026\)](#) fused TF-IDF lexical features, BioBERT-NER entities represented through UMLS-linked graphs, and PubMedBERT contextual embeddings using Canonical Correlation Analysis for cancer-type classification, providing a multi-view biomedical representation.

[Laxminarayana and Shaik \(2026\)](#) combined transformer-derived contextual and complementary semantic information for clinical and medical sentiment classification. This semantic-context fusion contributes to Tier 1, while confidence calibration contributes separately to Tier 2.

[Baqraf and Keikhosrokiani \(2026\)](#) investigated PCA- and ICA-transformed Arabic BERT representations, medical-domain pre-training, and an entropy-informed loss for health-webpage quality classification. Representation transformation and domain

adaptation support Tier 1, while explicit uncertainty and calibration evaluation contributes separately to Tier 2.

Lee et al. (2025) proposed Entity-enhanced BERT (E-BERT), integrating medical entities through entity embeddings and gated entity-aware self-attention for medical-specialty prediction, thereby explicitly incorporating medical entities and their contextual relationships.

Collectively, these approaches enrich representations through biomedical concepts, domain knowledge, feature transformation, semantic-context fusion, and entity-aware modelling, although most evaluate reliability primarily through classification performance.

4.2.3 Multi-Task and Hierarchical Representation

Durán-Silva et al. (2024) evaluated general and biomedical language models on the BATRACIO research-phase classification task, combining domain-specific pre-training with preprocessing and class-imbalance handling to distinguish semantically adjacent categories.

Zhang et al. (2023b) combined separate BERT-based title and abstract representations with TextRank keywords and TextRNN contextual information for fine-grained classification across 62 cancer labels.

Tao et al. (2026) proposed multi-scale feature fusion combining local contextual patterns, pooled representations, and global semantic information from pretrained language models.

Chen and Du (2025) combined a Self-Attentive Adversarial Augmentation Network with Disease-Aware Multi-Task BERT. The auxiliary disease co-occurrence task strengthens disease-aware representations for Tier 1, while GAN-based minority-class augmentation contributes separately to Tier 4.

Kodati and Dasari (2025) integrated Med-RoBERTa embeddings with clinical text and structured EHR information using BiGRU, intra-attention, multi-head attention, and cross-attention, combining contextual and multimodal evidence before classification.

Lakshitha B. et al. (2026) augmented RoBERTa with sequential attention for multiclass emotional-health classification, using the attention mechanism to integrate informative evidence across the input sequence.

These methods therefore strengthen complementary, hierarchical, task-specific, or multimodal evidence rather than directly estimating prediction-level uncertainty.

4.2.4 Clinical-Domain and Task-Specific Representation Approaches

Nowak et al. (2023) compared medically pre-trained and locally radiology-pretrained transformers for multilabel classification of German chest X-ray reports. Local domain pre-training supports Tier 1, while silver, gold, and hybrid label strategies contribute separately to Tier 4.

Bressem et al. (2020) developed German radiology-specific BERT models using large-scale radiology pre-training and a domain-specific WordPiece vocabulary,

reducing fragmentation of specialised terminology; report pre-selection contributes separately to Tier 4.

Ferdowsi et al. (2023) combined transformer-derived clinical-trial text representations with graph-based protocol structure for trial-risk prediction. This textual-structural fusion supports Tier 1, while retrospective ternary risk-label construction contributes separately to Tier 4.

Naseem and Catal (2023) addressed long MIMIC-III discharge summaries by segmenting documents, encoding segments independently with BioBERT, and aggregating them through attention.

Can and Yücedağ (2026) evaluated long-context transformers for user-level depression-signal detection. Their strongest configuration combined a 4096-token Longformer with attention pooling and gated fusion of behavioural, linguistic, and topic metadata.

Sinha et al. (2024) compared general and biomedical pretrained models across six English and French biomedical classification datasets. Domain-specific representation contributes to Tier 1, while DropConnect, Monte Carlo Dropout, variational inference, and predictive-reliability evaluation contribute separately to Tier 2.

Flores and Verschae (2025) constructed a Bag-of-Regular-Expressions representation for Spanish biomedical text classification using automatically generated regular expressions as interpretable task-specific features; maximum-entropy active learning and its stopping criterion contribute separately to Tier 4.

Collectively, these approaches address specialised terminology, long documents, structural complexity, multilingual settings, and application-specific evidence through domain adaptation, aggregation, feature fusion, and task-specific representation.

Taken together, the probabilistic class-conditional approaches proposed by Manage et al. (2026b) and Manage et al. (2026a) explicitly encode statistical associations between textual evidence and class labels. Knowledge-enriched methods incorporate biomedical concepts or entities (Saqib and Mustafa 2025; Rostam and Kertész 2026; Alnajjar et al. 2026; Lee et al. 2025). Other approaches use domain-specific pre-training, feature fusion, separate textual encoders, multi-task learning, structural modelling, attention, long-context encoding, metadata integration, or regular-expression-derived features (Baqraf and Keikhosrokiani 2026; Durán-Silva et al. 2024; Zhang et al. 2023b; Chen and Du 2025; Kodati and Dasari 2025; Nowak et al. 2023; Ferdowsi et al. 2023; Can and Yücedağ 2026; Sinha et al. 2024; Flores and Verschae 2025). These methods primarily strengthen the quality or class-discriminateness of evidence available to the classifier; classification gains alone do not establish calibration, uncertainty discrimination, or robustness under distribution shift. Tier 1 therefore represents representation-oriented reliability support rather than direct prediction-level uncertainty estimation.

4.3 Tier 2: Uncertainty Estimation Methods ($n = 5$)

Tier 2 comprises methods that explicitly quantify or evaluate prediction-level uncertainty, confidence, calibration, or model-output reliability. Unlike Tier 1, it evaluates the reliability of resulting predictions rather than the evidence representation itself.

Tier 2 accounts for 5/45 (11.1%) of all taxonomy assignments; confidence intervals or significance tests applied only to aggregate performance were not treated as prediction-level uncertainty.

4.3.1 Calibration and Predictive-Reliability Evaluation

[Baqraf and Keikhosrokiani \(2026\)](#) evaluated Arabic health-webpage classification using confidence, entropy, and class-conditional calibration measures. This explicit reliability assessment supports Tier 2, while PCA, ICA, and medical-domain pre-training contribute separately to Tier 1.

[Laxminarayana and Shaik \(2026\)](#) evaluated confidence reliability using expected calibration error, contributing to Tier 2; their semantic-context representation fusion contributes separately to Tier 1.

[Sinha et al. \(2024\)](#) compared frequentist and Bayesian uncertainty-aware classifiers across six English and French biomedical datasets. DropConnect, Monte Carlo Dropout, and variational inference were evaluated using entropy, Brier score, expected and static calibration error, negative log-likelihood, and coverage. Their uncertainty modelling supports Tier 2, while domain-specific representation comparison contributes to Tier 1.

Together, these studies demonstrate calibration-oriented evaluation using different uncertainty constructs and reliability measures.

4.3.2 Uncertainty Estimation Linked to Selective Classification

[Peluso et al. \(2024\)](#) evaluated post-hoc uncertainty scores for pathology-report classification to retain predictions satisfying a specified accuracy requirement while minimising rejection. The uncertainty scores contribute to Tier 2, while their use for automatic acceptance or manual assessment contributes separately to Tier 3.

[Šuster et al. \(2023\)](#) evaluated EvidenceGRADER and RobotReviewer using expected calibration error, Brier scores, reliability diagrams, and risk-coverage analysis across medical areas. Their calibration evaluation supports Tier 2, while confidence-thresholded selective classification contributes to Tier 3.

These studies connect uncertainty estimation with operational outcomes such as coverage, prediction risk, and rejection.

Overall, Tier 2 uses heterogeneous uncertainty constructs and evaluation protocols. [Baqraf and Keikhosrokiani \(2026\)](#) evaluate entropy and class-conditional calibration; [Laxminarayana and Shaik \(2026\)](#) evaluate confidence calibration; [Sinha et al. \(2024\)](#) compare frequentist and Bayesian approaches using multiple reliability measures; and [Peluso et al. \(2024\)](#) and [Šuster et al. \(2023\)](#) connect uncertainty or confidence assessment with selective classification. The absence of a common uncertainty construct and evaluation protocol limits direct comparison across models, datasets, and medical domains.

4.4 Tier 3: Uncertainty-Aware Decision Making ($n = 7$)

Tier 3 comprises methods that use uncertainty, confidence, prediction stability, or model agreement to determine an inference-time action, including acceptance, routing,

fallback, continued computation, rejection, human review, or abstention. Unlike Tier 2, which estimates or evaluates prediction reliability, Tier 3 requires the reliability-related signal to influence what the system does next. This tier accounts for 7/45 (15.6%) of all taxonomy assignments.

4.4.1 Confidence-Based Routing and Human Review

Manage et al. (2026a) used average probability difference (AvgDiff) to separate confident and uncertain medical-text instances. Confident instances follow the standard transformer pathway, whereas uncertain cases are routed to a WCPE-enhanced model. Confidence therefore determines the classification pathway for Tier 3, while WCPE contributes separately to Tier 1.

Flores and Verschae (2024) used a validation-derived probability threshold for Spanish clinical-text classification. Lower-confidence cases invoke automatically generated regular-expression rules, with the original classifier used as fallback when no suitable rule applies. Prediction confidence therefore determines which mechanism produces the final classification.

Purcaru et al. (2026) developed a confidence-based human-in-the-loop workflow for neurology outpatient-note classification. High-confidence predictions were accepted automatically, whereas lower-confidence cases were manually coded. Confidence therefore controls automatic acceptance versus human review, while subsequent use of these cases for targeted annotation contributes separately to Tier 4.

Together, these methods use confidence to route cases among standard, enhanced, rule-based, or human-review pathways.

4.4.2 Selective Classification, Early Exiting, and Abstention

Peluso et al. (2024) used post-hoc uncertainty scores for selective pathology-report classification. Predictions meeting a specified accuracy requirement were retained automatically, whereas less reliable cases were rejected for manual assessment. This acceptance-rejection mechanism supports Tier 3, while uncertainty-score development and evaluation contribute separately to Tier 2.

Šuster et al. (2023) used confidence thresholds and risk-coverage analysis to determine which EvidenceGRADER and RobotReviewer predictions could be retained automatically. Their confidence-based acceptance criterion supports Tier 3, while calibration and prediction-reliability analyses contribute separately to Tier 2.

Zhan et al. (2026) proposed PEER, a patience-based early-exiting and rejection framework. Layer-wise prediction stability determines whether biomedical-text samples exit early, continue through additional transformer layers, or are rejected.

Greatrix et al. (2026) used heterogeneous unanimous ensembles combining zero-shot GPT-4o with task-specific models for pathology-report classification. Predictions were accepted only under unanimous agreement; disagreement triggered abstention and expert review. Agreement therefore acts as the Tier 3 decision criterion, while creation of a high-reliability labelled subset contributes separately to Tier 4.

These methods extend uncertainty-aware decisions beyond confidence thresholds through post-hoc uncertainty scores, risk-coverage criteria, layer-wise stability, and model agreement.

Overall, [Manage et al. \(2026a\)](#) and [Flores and Verschae \(2024\)](#) route uncertain cases to alternative classification pathways; [Purcaru et al. \(2026\)](#) use confidence to separate automatic coding from human review; [Peluso et al. \(2024\)](#) and [Šuster et al. \(2023\)](#) use selective-classification criteria; [Zhan et al. \(2026\)](#) use prediction stability to control early exiting or rejection; and [Greatrix et al. \(2026\)](#) require unanimous agreement for automatic acceptance.

However, decision rules and evaluation procedures remain task-specific, using different probability thresholds, accuracy requirements, risk-coverage criteria, patience conditions, rejection rules, and agreement requirements. Reported outcomes also vary across accuracy, coverage, rejection, computational efficiency, and human-review burden. The evidence therefore does not establish a common decision rule or evaluation protocol for uncertainty-aware inference across medical text-classification settings.

4.5 Tier 4: Training-Data Reliability ($n = 12$)

Tier 4 comprises methods that control the construction, selection, labelling, filtering, weighting, or influence of training evidence. These approaches address reliability problems involving synthetic data, weak or automatically generated labels, silver-standard annotations, retrospective labels, and selectively acquired examples. Tier 4 accounts for 12/45 (26.7%) of all taxonomy assignments.

4.5.1 Synthetic and Augmented Training Data

[Chen and Du \(2025\)](#) combined a Self-Attentive Adversarial Augmentation Network with Disease-Aware Multi-Task BERT. The augmentation network generates minority-class examples for imbalanced medical-text classification, contributing to Tier 4, while the disease-aware representation mechanism contributes separately to Tier 1.

[Inan \(2026\)](#) used LLaMA 3.1 to generate synthetically labelled sentences for Hallmarks of Cancer classification. BioLinkBERT-based retrieval and maximal marginal relevance selected relevant and diverse few-shot examples, and generated data were used alone or with human-annotated examples to fine-tune BioLinkBERT-Large. The Tier 4 contribution is the construction and selection of synthetic labelled training evidence.

[Duan et al. \(2025\)](#) proposed MedAS, a GPT-4-assisted framework for synthetic electronic medical records. Knowledge-based checks identified clinically inconsistent or diagnostically invalid records, while noise-tolerant training assigned lower-confidence soft targets until model confidence supported promotion to a more reliable training set. This controls the filtering and influence of potentially unreliable synthetic evidence.

Together, these studies range from direct augmentation to pipelines that generate, select, verify, down-weight, or progressively admit synthetic examples before training.

4.5.2 Label Construction, Weak Supervision, and Automated Labelling

Nowak et al. (2023) compared rule-generated silver labels, manually curated gold labels, and a hybrid silver-to-gold strategy for German chest X-ray report classification. Evaluation of annotation source and quality supports Tier 4, while local radiology-specific pre-training contributes separately to Tier 1.

Bressem et al. (2020) introduced automated pre-selection of German chest-radiograph reports containing insufficient information for subsequent multilabel finding extraction. This controls which reports are sufficiently informative for label construction, while radiology-specific pre-training and vocabulary construction contribute separately to Tier 1.

Ferdowsi et al. (2023) constructed retrospective low-, medium-, and high-risk labels for clinical-trial protocols from historical indicators including trial status, enrolment changes, duration divergence, protocol amendments, and primary-outcome changes. Label construction contributes to Tier 4, while textual and graph-based representation learning contributes separately to Tier 1.

Khandokar et al. (2024) generated initial disease-related weak labels from uncertainty-related lexical information and semantic proximity and refined them using multiple classifiers before training multilingual BERT. The contribution is therefore weak-label generation and refinement rather than prediction-level uncertainty estimation.

Cardamone et al. (2025) compared GPT-4-generated classifications of unstructured electronic health record terms with consensus decisions from clinical experts, thereby evaluating the reliability of automatically generated labels intended to support downstream modelling.

Greatrix et al. (2026) used heterogeneous unanimous ensembles combining zero-shot GPT-4o with task-specific models to construct automatically labelled pathology-report subsets. Labels were retained only under unanimous agreement, with disagreements referred for expert review. Agreement therefore controls label reliability for Tier 4, while abstention contributes separately to Tier 3.

Collectively, these studies address training-data reliability through report selection, label-source comparison, retrospective label construction, weak supervision, expert validation, and model agreement.

4.5.3 Active Learning and Targeted Annotation

Purcaru et al. (2026) manually coded low-confidence neurology outpatient notes and subsequently incorporated them into the training data, contributing to Tier 4 through targeted annotation and training-data enrichment. Confidence-based automatic acceptance versus human review contributes separately to Tier 3.

Flores et al. (2021) extended the CREGEX biomedical classifier with active learning, selecting unlabelled examples for expert annotation using regular-expression precision and Smith-Waterman similarity and applying a variance-based stopping criterion. The method therefore controls training-data acquisition through informativeness.

Flores and Verschae (2025) used maximum-entropy sampling to select informative Spanish biomedical texts and variance of entropy as an annotation stopping criterion. Uncertainty therefore guides training-data acquisition and annotation effort, while the Bag-of-Regular-Expressions representation contributes separately to Tier 1.

Overall, Tier 4 spans synthetic-data generation and control in Chen and Du (2025), Inan (2026), and Duan et al. (2025); report and label reliability in Bressem et al. (2020), Nowak et al. (2023), Ferdowsi et al. (2023), Khandokar et al. (2024), Cardamone et al. (2025), and Greatrix et al. (2026); and targeted annotation in Purcaru et al. (2026), Flores et al. (2021), and Flores and Verschae (2025).

Despite this breadth, Tier 4 studies do not use a shared definition or validation procedure for training-data reliability. They evaluate different evidence types, including synthetic records, weak labels, silver and gold annotations, retrospective labels, automatically generated labels, and actively selected examples, using combinations of downstream classification performance, expert agreement, confidence thresholds, annotation efficiency, and internal filtering criteria. Consequently, the evidence does not establish a common criterion for determining when generated, weakly labelled, or selectively annotated evidence is sufficiently reliable to influence training.

5 Discussion

5.1 Taxonomy Summary and Comparative Analysis

The four-tier taxonomy distinguishes reliability interventions at the representation, prediction, decision, and training-data levels. The 32 primary studies produced 45 assignments, with 13 studies contributing at two functional intervention points.

The distribution reveals an important methodological imbalance. Tier 1 is dominated by methods that strengthen class-relevant or domain-specific representations, but improved classification performance alone does not establish calibrated or reliable predictions. Tier 2 directly evaluates predictive reliability, yet its relatively small evidence base and heterogeneous uncertainty measures limit comparison across settings.

Tier 3 demonstrates that uncertainty, confidence, prediction stability, and model agreement can support routing, selective acceptance, human review, early exiting, rejection, and abstention. However, these decision mechanisms remain largely dependent on task-specific rules whose transferability has not been established. Tier 4 further shows that reliability depends on training evidence, but synthetic-data, label-quality, weak-supervision, and targeted-annotation approaches lack a shared validation criterion.

Overall, reliability in medical text classification operates across interconnected intervention points rather than at prediction time alone. The principal methodological gaps are the limited use of comparable prediction-level reliability evaluation and the scarcity of externally validated uncertainty-aware decision procedures.

The study-level evidence underlying this synthesis is presented in Table 3.

Table 3: Comparative analysis of the 32 unique primary studies included in the systematic synthesis. Each study is listed once; multiple assignments are shown together in the Tier column. The 13 cross-tier studies yield 45 taxonomy assignments. T1 = feature representation; T2 = uncertainty estimation; T3 = uncertainty-aware decision making; and T4 = training-data reliability.

No.	Study	Domain / Data	Backbone	Reliability-related contribution	Tier	Principal reported finding
1	Manage et al. (2026b)	PubMed 20K and Ohsumed	BERT with Term-Class Probability Embedding (TCPE)	Smoothed term-class conditional probabilities and TF-IDF weighting inject class-discriminative lexical evidence into BERT document representations.	T1	TCPE improved class separation and classification performance without substantial additional computational cost.
2	Manage et al. (2026a)	PubMed, Ohsumed, and Drug Review	BERT, BioBERT, and PubMedBERT with Word-Class Probabilistic Embedding (WCPE)	WCPE strengthens document representations, while AvgDiff separates confident and uncertain documents and determines whether WCPE-enhanced classification is applied.	T1/T3	The two-stage framework improved classification of uncertain and boundary-case documents by routing them to the WCPE-enhanced model.
3	Saqib and Mustafa (2025)	PMC-Patients clinical notes	PubMedBERT with UMLS Concept Unique Identifiers	Raw and named-entity-recognition-filtered UMLS concepts are appended to clinical-text inputs to provide structured biomedical knowledge.	T1	Filtered UMLS augmentation improved diagnosis classification, particularly for minority classes, whereas unfiltered augmentation introduced noise and reduced performance.
4	Rostam and Kertész (2026)	PubMed abstracts for EQ-5D study detection	BERT, SciBERT, and BioBERT with scispaCy entity enrichment and attention-based multiple-instance learning	Biomedical entities enrich sentence representations, while multiple-instance learning aggregates sentence-level evidence for study-level classification.	T1	The method improved automated screening of EQ-5D studies, with the best study-level model achieving an F1-score of approximately 0.82 and high recall.
5	Alnajjar et al. (2026)	Biomedical publications concerning thyroid, colon, and lung cancer	TF-IDF features, BioBERT-NER, UMLS-linked entity graphs, PubMedBERT, and Canonical Correlation Analysis	Lexical, biomedical entity-graph, and contextual representations are combined through multi-view feature fusion.	T1	The hybrid representation improved three-class cancer-document classification and achieved approximately 0.91 accuracy.
6	Laxminarayana and Shaik (2026)	MIMIC-III and i2b2 clinical sentiment corpora	BioClinicalBERT with hierarchical contextual modelling and gated semantic-context fusion	Transformer-derived semantic representations are combined with discourse-level contextual representations, while Expected Calibration Error evaluates confidence reliability.	T1/T2	The complete framework achieved a macro-F1 score of 0.904 and reduced Expected Calibration Error to 0.041, with particularly strong improvement for uncertain and negative sentiment classes.

Continued on the next page

Table 3 continued from the previous page

No.	Study	Domain / Data	Backbone	Reliability-related contribution	Tier	Principal reported finding
7	Baqrāf and Keikhosrokiani (2026)	Arabic online health-information webpages	Arabic BERT and MAR-BERT with PCA, ICA, and an entropy-informed loss	PCA and ICA transform BERT representations, while entropy, confidence, and class-conditional expected calibration error evaluate predictive certainty and calibration.	T1/T2	The PCA-based model achieved approximately 94.7% accuracy, while the entropy-informed model produced more balanced class-conditional calibration.
8	Lee et al. (2025)	Korean clinical questionnaire data for medical-specialty prediction	Entity-enhanced BERT with medical NER, entity embeddings, and gated entity-aware self-attention	Symptom, anatomical-location, and disease entities enrich contextual representations and model relationships between medical entities.	T1	Entity-enhanced representations improved four-class medical-specialty prediction, including performance on longer clinical inputs.
9	Durán-Silva et al. (2024)	BATRACIO biomedical-publication dataset	BERT, BioBERT, SciBERT, PubMedBERT, and LinkBioBERT	General and biomedical domain-specific representations are compared using titles and abstracts, preprocessing, acronym expansion, weighted loss, and class-balancing strategies.	T1	Domain-specific pretrained language models effectively classified publications as basic, translational, clinical, public-health, or other research, although performance varied across adjacent and imbalanced categories.
10	Zhang et al. (2023b)	70,599 cancer publications from the Dimensions database	Dual-branch BERT with TextRank and TextRNN	Separate title and abstract representations are combined with abstract keywords and sequential contextual information for multilabel classification.	T1	The BERT-TextRNN model supported publication-level classification across 62 cancer labels and achieved an F1-score of 90.34%.
11	Tao et al. (2026)	CHIP-CTC, KUAKE-QIC, CMeEE, and CHIP-MDCFNPC	RoBERTa-wwm-ext with multi-scale convolution, double pooling, and global semantic integration	Local features at multiple receptive-field scales are combined with pooled contextual information and global semantic features.	T1	The multi-scale feature-fusion framework improved medical short-text classification across several benchmark datasets, while ablation experiments supported the contribution of its three principal modules.
12	Chen and Du (2025)	Chinese cardiovascular clinical records and CCKS 2017 medical text	Self-Attentive Adversarial Augmentation Network and Disease-Aware Multi-Task BERT	A sparse-attention generative adversarial network creates minority-class training samples, while an auxiliary disease co-occurrence task strengthens disease-aware representations.	T1/T4	The combined framework improved rare-class recognition and classification under class imbalance, achieving an overall F1-score of approximately 0.91 in the reported comparison.
13	Kodati and Dasari (2025)	MIMIC-III clinical notes and linked MIMIC-III/MIMIC-IV EHR information	Med-RoBERTa with Hybrid-IMHA-BiGRU	Clinical text and structured EHR features are integrated using dual-channel processing, BiGRU, intra-attention, multi-head attention, and cross-attention.	T1	The hybrid representation identified diseases associated with higher mortality by capturing contextual, long-range, and variable-length EHR evidence.

Continued on the next page

Table 3 continued from the previous page

No.	Study	Domain / Data	Backbone	Reliability-related contribution	Tier	Principal reported finding
14	Lakshitha B. et al. (2026)	53,043 emotional-health narratives across seven classes	RoBERTa-BiLSTM with multi-head attention	Contextual transformer embeddings are combined with sequential modelling and attention-based integration of emotionally informative textual evidence.	T1	The hybrid model achieved 79.0% accuracy, a weighted F1-score of 0.79, and a macro-F1 score of 0.73, outperforming transformer-only and conventional baselines.
15	Nowak et al. (2023)	German free-text chest X-ray reports	Public German medical BERT and locally masked-language-model-pretrained BERT	Local radiology pre-training strengthens domain-specific representations, while silver, gold, and hybrid silver-to-gold labels control the quality of training supervision.	T1/T4	The study supported multilabel structuring of radiology reports and demonstrated that performance depends on both domain pre-training and annotation quality.
16	Bressem et al. (2020)	German chest-radiograph reports	Radiology-specific RAD-BERT and FS-BERT models with a BERT-based report pre-selection stage	Domain-specific pre-training and vocabulary construction strengthen radiology-text representations, while pre-selection excludes reports that are insufficient for reliable label extraction.	T1/T4	The report pre-selection models identified unevaluable reports with F1-scores of 0.97–0.98, supporting more reliable downstream extraction of radiological findings.
17	Ferdowsi et al. (2023)	ClinicalTrials.gov interventional trial protocols	Sampled transformer language model, graph neural network, and probability-level ensemble	Textual protocol representations are combined with graph-based section structure, while retrospective protocol histories are used to construct low-, medium-, and high-risk labels.	T1/T4	The ensemble predicted clinical-trial risk categories and achieved a reported AUROC of 0.8453.
18	Naseem and Catal (2023)	MIMIC-III discharge summaries	BioBERT with document segmentation and attention-based embedding aggregation	Long discharge summaries are divided into segments, independently encoded with BioBERT, and combined through attention-weighted aggregation.	T1	The framework achieved an AUROC of 0.69 for patient readmission prioritisation and demonstrated the feasibility of attention-based aggregation for long clinical documents.
19	Can and Yücedağ (2026)	User-level social-media timelines for depression-signal detection	BERT-family models, Longformer, and BigBird with gated metadata fusion	Long-context text representations are combined with behavioural, linguistic, stylistic, and topic-related metadata.	T1	The 4096-token Longformer with meta-data achieved an F1-score of 87.15% and an AUROC of 0.9667 for user-level depression-signal detection.
20	Sinha et al. (2024)	Six English and French biomedical text-classification datasets	BERT, BioBERT, CamemBERT, and CamemBERT-bio with frequentist and Bayesian classifier variants	Domain-specific representations are compared with DropConnect, Monte Carlo Dropout, and variational-inference approaches using entropy and calibration measures.	T1/T2	Domain-specific models generally improved classification, while uncertainty-aware variants often improved calibration; the relative benefit remained dependent on the dataset and task.

Continued on the next page

Table 3 continued from the previous page

No.	Study	Domain / Data	Backbone	Reliability-related contribution	Tier	Principal reported finding
21	Flores and Verschae (2025)	Spanish clinical texts concerning obesity and smoking	Bag-of-Regular-Expressions with Naive Bayes, Random Forest, XGBoost, and SVM ^a	Automatically generated regular expressions form an interpretable feature space, while maximum-entropy sampling and an entropy-variance stopping rule guide annotation.	T1/T4	The method provided interpretable clinical-text features and reduced the amount of labelled training data required through active learning.
22	Peluso et al. (2024)	Electronic pathology reports for cancer-registry classification	Deep learning text classifier with post-hoc uncertainty quantification ^a	Post-hoc uncertainty scores estimate prediction reliability and support selective classification under specified accuracy and rejection requirements.	T2/T3	The uncertainty methods identified predictions that could be retained automatically while rejecting less reliable reports for manual assessment.
23	Šuster et al. (2023)	Medical-evidence quality and risk-of-bias assessment	EvidenceGRADER and RobotReviewer	Expected calibration error, Brier scores, reliability diagrams, and risk-coverage curves evaluate calibration and confidence-thresholded selective classification.	T2/T3	Calibration and selective-prediction reliability varied substantially across medical areas and assessment tasks; reducing coverage lowered prediction risk for several tasks.
24	Zhan et al. (2026)	MIMIC-III and other biomedical text-classification datasets	Transformer backbones with Patience-Based Early Exiting with Rejection (PEER)	Layer-wise prediction consistency determines whether a prediction exits early, continues through additional layers, or is rejected rather than automatically classified.	T3	On MIMIC-III, PEER achieved 90.73% accuracy while rejecting 2.79% of uncertain samples and also supported substantial inference-speed improvements.
25	Flores and Verschae (2024)	Spanish clinical texts concerning smoking and obesity status	SVM, Random Forest, Naive Bayes, and BERT with automatically generated regular expressions	A validation-derived confidence threshold determines whether the classifier output is accepted or a precision-ranked regular-expression rule is applied.	T3	Confidence-based routing combined statistical or transformer predictions with an interpretable rule-based fallback for lower-confidence cases.
26	Purcaru et al. (2026)	Neurology outpatient care notes	Logistic regression, SVM, BiLSTM, DistilBERT, and RoBERTa with confidence-based human review	Confidence controls automatic acceptance or manual coding, while lower-confidence cases are selected for targeted annotation and training-data enrichment.	T3/T4	At a confidence threshold of 0.85, DistilBERT achieved 96% accuracy while automatically covering 64% of the evaluated cases.
27	Greatrix et al. (2026)	TCGA pathology reports for primary-diagnosis classification	Zero-shot GPT-4o with DistilBERT, SVM, and dense-neural-network unanimous ensembles	A prediction is accepted only when all constituent models agree; disagreement causes abstention and identifies cases for expert review or annotation.	T3/T4	The GPT-4o-DistilBERT unanimous ensemble achieved 95.5% accuracy on the 45.5% of test reports for which the models agreed, automatically identifying a high-reliability subset.
28	Khandokar et al. (2024)	Disease-related clinical reports	Lexical weak-labelling, semantic similarity, classifier ensemble, and multilingual BERT	Initial weak labels derived from uncertainty-related lexical information and semantic proximity are refined using multiple classifiers before training the final classifier.	T4	The pipeline used lexical and ensemble evidence to refine automatically generated training labels before multilingual BERT classification.

Continued on the next page

Table 3 continued from the previous page

No.	Study	Domain / Data	Backbone	Reliability-related contribution	Tier	Principal reported finding
29	Cardamone et al. (2025)	Unstructured EHR terms for mental- and physical-health prediction features	GPT-4 classification with clinical-expert consensus	Automatically generated feature-category labels are evaluated against consensus decisions from clinical experts to assess their suitability for downstream modelling.	T4	The study quantified the reliability of GPT-4-generated clinical feature labels and showed that agreement with experts varied across feature categories.
30	Flores et al. (2021)	Spanish clinical texts concerning smoking and obesity	CREGEX active learning with automatically generated regular expressions ^a	Regular-expression precision and Smith-Waterman sequence-similarity scores select informative unlabelled examples, while a variance-based rule determines when annotation should stop.	T4	The active-learning procedure reduced expert annotation requirements while maintaining biomedical text-classification performance.
31	Inan (2026) Inan (2026)	Hallmarks of Cancer corpus	LLaMA 3.1 synthetic-data generation with BioLinkBERT-Large fine-tuning	Retrieval-guided prompting and diversity-aware selection are used to generate labelled biomedical sentences that supplement the human-annotated training data.	T4	Augmenting the original data with LLM-generated samples achieved an average micro-F1 score of 85.09, outperforming fine-tuning using the original human-annotated data alone.
32	Duan et al. (2025)	Synthetic electronic medical records for multi-disease prediction	GPT-4-assisted MedAS with BERT-based masked-language modelling and noise-tolerant training	Knowledge-based checks filter diagnostically inconsistent synthetic records, while lower-confidence soft targets limit their influence until sufficient model confidence is reached.	T4	The framework improved the reliability and controlled use of synthetic EMR data by filtering inconsistent records and reducing the influence of potentially unreliable examples.

^a The principal proposed method did not use a transformer. The study was retained because it addressed a directly comparable reliability function within medical text classification.

The patterns identified in this synthesis and comparative analysis provide the basis for the research gaps discussed in the following subsections.

5.2 Research Gap 1: Limited Investigation of Explicit Class-Conditional Probabilistic Representation

Although Tier 1 is the largest category, most methods strengthen representations through biomedical knowledge, domain-specific pre-training, contextual encoding, feature fusion, attention, long-context modelling, structural information, or task-specific features rather than explicitly representing class-conditional evidence.

Within the reviewed studies, explicit term- or word-class probabilistic representation was concentrated in the TCPE and WCPE approaches proposed by [Manage et al. \(2026b\)](#) and [Manage et al. \(2026a\)](#). Evidence for these approaches remains limited across datasets, transformer backbones, languages, and application settings. Future research should evaluate their effects on classification, calibration, uncertainty discrimination, selective-classification behaviour, distribution-shift robustness, and rare or overlapping classes, while distinguishing gains from those attributable to domain-specific pre-training or increased representation capacity.

5.3 Research Gap 2: Limited Cross-Setting Validation of Uncertainty-Aware Decision Rules

Tier 3 studies operationalise prediction reliability in several ways: [Manage et al. \(2026a\)](#) route uncertain documents to an enhanced classifier; [Flores and Verschae \(2024\)](#) use confidence to select a classifier or rule-based fallback; [Purcaru et al. \(2026\)](#) separate automatic coding from manual assessment; [Peluso et al. \(2024\)](#) and [Šuster et al. \(2023\)](#) apply selective classification; [Zhan et al. \(2026\)](#) use prediction stability for early exiting or rejection; and [Greatrix et al. \(2026\)](#) require unanimous agreement for automatic acceptance.

However, these policies rely on heterogeneous, task-specific thresholds, risk-coverage criteria, patience conditions, rejection rules, or agreement requirements. Evidence remains limited that such rules transfer reliably across datasets, institutions, patient populations, classes, architectures, or distribution shifts, and no common procedure balances coverage, prediction risk, computational cost, and human-review burden.

Future work should evaluate class-specific risk, calibration, coverage, rejection, false-negative risk, computational cost, and human-review burden under temporal, institutional, linguistic, prevalence, and broader distribution shifts. Adaptive thresholding, risk-controlling prediction, conformal prediction, cost-sensitive deferral, and dynamic early-exit policies should also be compared with fixed-threshold, routing, selective-classification, and agreement-based approaches.

5.4 Research Gap 3: Lack of a Shared Validation Framework for Training-Data Reliability

The third gap concerns how training-example and label reliability are defined and validated. Tier 4 spans report selection, synthetic-data generation and filtering, label construction and refinement, automated labelling, active learning, and targeted annotation.

These functions are operationalised differently. [Bressem et al. \(2020\)](#) use report pre-selection; [Chen and Du \(2025\)](#) generate minority-class examples; [Inan \(2026\)](#) uses retrieval-guided, diversity-aware generation; and [Duan et al. \(2025\)](#) combine knowledge-based filtering with noise-tolerant control of synthetic evidence. [Nowak et al. \(2023\)](#) compare silver, gold, and hybrid labels; [Ferdowsi et al. \(2023\)](#) construct retrospective risk labels; [Khandokar et al. \(2024\)](#) refine weak labels; [Cardamone et al. \(2025\)](#) compare automated labels with expert consensus; [Greatrix et al. \(2026\)](#) use unanimous agreement; and [Purcaru et al. \(2026\)](#), [Flores et al. \(2021\)](#), and [Flores and Verschae \(2025\)](#) prioritise examples for annotation using confidence or informativeness.

However, there is no shared definition of reliable training evidence. Reliability is inferred from downstream performance, confidence or agreement, expert judgement, consistency checks, filtering criteria, or annotation efficiency. These measures are not interchangeable: downstream performance does not establish clinical validity, and high confidence or agreement does not guarantee factual correctness, label validity, or representativeness.

Future work should separately assess label correctness, clinical and factual consistency, linguistic quality, annotation agreement, distributional compatibility, and downstream effects. Shared reporting standards and benchmarks are needed to compare synthetic, weakly labelled, automatically labelled, and selectively acquired training evidence across medical-text classification settings.

5.5 Practical Implications

The reviewed evidence does not support using raw predictive confidence as the sole basis for autonomous medical text-classification decisions. Systems intended for clinical or operational use should be evaluated using calibration, class-specific performance, risk-coverage behaviour, rejection or deferral behaviour, and human-review requirements. Human oversight should be maintained until the relevant uncertainty-aware decision mechanism has been externally validated across appropriate datasets, institutions, patient populations, and deployment conditions.

5.6 Future Research Directions

Five priorities emerge from the synthesis. First, predictive reliability should be evaluated consistently alongside conventional classification performance using calibration measures, Brier score or negative log-likelihood, class-specific reliability, uncertainty discrimination, and risk-coverage behaviour, with calculation procedures clearly reported.

Second, uncertainty-aware decision policies should be evaluated as risk-sensitive mechanisms under temporal, institutional, linguistic, prevalence, and broader distribution shifts. Relevant outcomes include class-specific risk, coverage, rejection, computational cost, and human-review burden. Adaptive thresholding, risk-controlling prediction, conformal prediction, cost-sensitive deferral, and dynamic early exiting warrant comparison with fixed decision rules.

Third, representation research should determine whether improved class discrimination also improves predictive reliability by evaluating probabilistic class-conditional, biomedical knowledge-enriched, and long-context representations using both classification and calibration outcomes across domains, languages, and transformer architectures.

Fourth, training-data reliability should be examined through synthetic-data verification, weak-label refinement, annotation disagreement, label-noise detection, class imbalance, distributional compatibility, filtering, weighting, soft labelling, and targeted annotation.

Fifth, generative large language models should have clearly defined and evaluated roles in uncertainty-aware pipelines, including training-data generation, label review, secondary assessment of uncertain cases, and heterogeneous ensembling. Their reliability risks should be evaluated alongside performance benefits using cross-institutional, multilingual, and prospective validation.

5.7 Limitations

This review has several limitations. First, the search strategy was deliberately transformer-centred and included terms such as BERT, BioBERT, and transformer. Although several principally non-transformer studies were retained for methodological comparison, some Bayesian, rule-based, or conventional machine-learning approaches addressing comparable reliability functions may not have been identified. Accordingly, the review should be interpreted as a focused synthesis of transformer-centred medical text classification rather than an exhaustive review of all uncertainty-aware NLP methods. Second, database and platform export limitations affected retrieval. Of the 5,895 records identified, 4,488 were successfully imported into Mendeley Reference Manager for deduplication and screening. The affected sources, numbers of records, and export constraints are reported transparently in the search methodology. Third, the search ended in May 2026, so studies published, indexed, or peer-reviewed thereafter were outside the review window. This is particularly relevant to rapidly developing generative-LLM and medical-NLP research. Fourth, study selection and initial taxonomy coding were led principally by the first author using predefined eligibility and classification rules, with borderline decisions discussed with the supervisory authors. This structured approach supported consistency in applying the review framework; however, because screening and coding were not fully duplicated by independent reviewers, some susceptibility to selection or classification bias remains, particularly for studies requiring functional interpretation across heterogeneous methods. Fifth, the review focused on functional methodological synthesis rather than formal comparative-effect estimation and therefore did not undertake study-level risk-of-bias, reporting-bias, or certainty-of-evidence grading. This approach enabled comparison

of heterogeneous reliability mechanisms across diverse medical-text classification settings; however, the findings should be interpreted as a methodological synthesis rather than as a formally graded comparison of study quality or effect strength.

Finally, the included studies differed substantially in datasets, languages, classification tasks, class distributions, architectures, evaluation measures, and definitions of uncertainty or reliability. Quantitative pooling was therefore not considered appropriate for the aims of this review, and direct comparison of reported performance across studies remains limited.

Despite these limitations, the review provides a transparent and structured functional synthesis distinguishing representation-, prediction-, decision-, and training-data-level reliability contributions in transformer-centred medical text classification.

6 Conclusion

This systematic methodological review synthesised 32 primary studies and organised their reliability-related contributions into a four-tier functional taxonomy covering feature representation, uncertainty estimation, uncertainty-aware decision making, and training-data reliability. Because 13 studies contributed substantively at two intervention points, the review identified 45 taxonomy assignments.

The synthesis shows that reliability in transformer-centred medical text classification is addressed most often through representation enhancement and training-data reliability, whereas explicit prediction-level uncertainty estimation remains less extensively investigated. Improvements in classification performance do not by themselves establish calibrated or trustworthy predictions; uncertainty-aware decision mechanisms remain largely task-specific; and training-data reliability approaches lack a shared validation framework.

Reliable medical text classification therefore requires coordinated attention to representation quality, prediction-level uncertainty, inference-time decision policies, and the reliability of training evidence. Future research should combine classification and calibration evaluation, validate uncertainty-aware decision rules across settings, and establish transparent criteria for assessing training-data reliability. The proposed four-tier taxonomy provides a functional framework for comparing these intervention points and guiding the development of more reliable transformer-centred medical text-classification systems.

A Study-Level Reasons for Full-Text Exclusion

Appendix Table 4 reports the full-text reports formally excluded from the primary synthesis and the corresponding study-level reasons for exclusion under the predefined eligibility criteria. The IC1–IC5 columns indicate whether each report satisfied the five inclusion criteria, while the Reason / FE Code column records the applicable formal exclusion criterion.

Table 4: Full-text reports excluded from the primary synthesis and reasons for exclusion.

No.	Title	Author(s)	Year	Decision	Reason / FE Code	IC1	IC2	IC3	IC4	IC5	Category
1	Mind the Ambiguity: Aleatoric Uncertainty Quantification in LLMs for Safe Medical Question Answering	Liu et al.	2026	EXCLUDE	FE4: Medical question answering/generation rather than eligible sample-level text classification.	N	Y	Y	Y	Y	Formal exclusion
2	EnergyRoute: Energy-Based Uncertainty Routing for Selective Retrieval and Large Language Model Assistance in the Hierarchical Classification of Biotechnology R&D Projects	Jang et al.	2026	EXCLUDE	FE2: Biotechnology R&D project classification is an administrative/scientific-document task, not medical or health text classification.	N	Y	Y	Y	Y	Formal exclusion
3	Mutual Learning and Adaptive Ensembling of Pre-Trained Language Models for Hierarchical and Flat Scientific Text Classification	Rostam, Z. R. K. and Kertész, G	2026	EXCLUDE	FE2: Primary task is general scientific-text classification rather than medical, clinical, biomedical, or health text classification.	N	Y	Y	Y	Y	Formal exclusion
4	Hierarchical Classification for Differential Diagnosis of Fever of Unknown Origin: A Multi-Task Learning Approach with Self-Adaptive Representation Sharing	Wang et al.	2026	EXCLUDE	FE3: Medical classification study without a substantive contribution to predictive uncertainty, uncertainty-aware decisions, representation reliability, or training-data reliability.	Y	Y	N	Y	Y	Formal exclusion

Continued on the next page

Table 4 (continued)

No.	Title	Author(s)	Year	Decision	Reason / FE Code	IC1	IC2	IC3	IC4	IC5	Category
5	A Hybrid Approach for Predicting Depression Using Deep Learning	Krishna et al.	2025	EXCLUDE	FE3: Performance-oriented hybrid depression classifier without a qualifying reliability, calibration, decision, or training-data-reliability mechanism.	Y	N	N	Y	Y	Formal exclusion
6	Granular-Ball Representation-Based Two-Stage Deep Learning Model for Text Classification	Qian et al.	2025	EXCLUDE	FE2: General-domain text classification without an eligible medical or health primary dataset.	N	Y	Y	Y	Y	Formal exclusion
7	An Uncertainty-Aware Decision Support System: Integrating Text Narratives and Conformal Prediction for Trustworthy Accident Code Classification	Kumar et al.	2025	EXCLUDE	FE2: Classifies industrial accident narratives in occupational/process safety rather than medical or health text.	N	Y	Y	Y	Y	Formal exclusion
8	Uncertainty Quantification for Machine Learning in Healthcare: A Survey	Lechuga Lopez et al.	2025	EXCLUDE	FE9: Survey article rather than an eligible primary experimental medical text-classification study.	N	N	Y	N	Y	Formal exclusion
9	Word-Sequence Entropy: Towards Uncertainty Estimation in Free-Form Medical Question Answering Applications and Beyond	Wang et al.	2025	EXCLUDE	FE4: Primary task is free-form medical question answering/generation rather than eligible sample-level text classification.	N	Y	Y	Y	Y	Formal exclusion
10	Three-Way Multi-Label Classification: A Review, a Framework, and New Challenges	Zhang et al.	2025	EXCLUDE	FE9: Review/framework article rather than an eligible primary medical text-classification experiment.	N	N	Y	N	Y	Formal exclusion
11	NLP-Powered Healthcare Insights: A Comparative Analysis for Multi-Labeling Classification with the MIMIC-CXR Dataset	Uslu et al.	2024	EXCLUDE	FE3: Standard comparative medical classification study without a substantive reliability-related methodological contribution.	Y	Y	N	Y	Y	Formal exclusion

Continued on the next page

Table 4 (continued)										
No.	Title	Author(s)	Year	Decision	Reason / FE Code	IC1	IC2	IC3	IC4	IC5Category
12	Iterative Multitask Learning and Inference from Seismic Images	Gao	2024	EXCLUDE	FE1: Image-based geophysical/seismic task with no substantive medical text-classification component.	N	N	N	Y	Y Formal exclusion
13	MUTUAL: Multi-Domain Sentiment Classification via Uncertainty Sampling	Katsarou et al.	2023	EXCLUDE	FE2: Uses non-medical sentiment datasets; the primary task lies outside the review domain.	N	Y	Y	Y	Y Formal exclusion
14	Robust Fake Review Detection Using Uncertainty-Aware LSTM and BERT	Zabeen et al.	2023	EXCLUDE	FE2: Fake-review detection is a non-medical text-classification task.	N	Y	Y	Y	Y Formal exclusion
15	Proximity-Informed Calibration for Deep Neural Networks	Xiong et al.	2023	EXCLUDE	FE2: General calibration study evaluated outside the eligible medical-text classification task space.	N	Y	Y	Y	Y Formal exclusion
16	More Sustainable Text Classification via Uncertainty Sampling and a Human-in-the-Loop	Andersen and Zukunft	2023	EXCLUDE	FE2: General-domain text classification and annotation workflow without a medical primary dataset.	N	Y	Y	Y	Y Formal exclusion
17	Towards More Reliable Text Classification on Edge Devices via a Human-in-the-Loop	Andersen and Zukunft	2022	EXCLUDE	FE2: General edge-device text classification without an eligible medical primary dataset.	N	Y	Y	Y	Y Formal exclusion
18	Improving Classification of Infrequent Cognitive Distortions: Domain-Specific Model vs. Data Augmentation	Ding et al.	2022	EXCLUDE	FE3: Domain-specific modelling and generic augmentation improve performance but do not provide a qualifying reliability mechanism under the final taxonomy.	Y	Y	N	Y	Y Formal exclusion
19	A Survey on Uncertainty Estimation in Deep Learning Classification Systems from a Bayesian Perspective	Mena et al.	2021	EXCLUDE	FE9: Survey article rather than a primary experimental medical text-classification study.	N	N	Y	N	Y Formal exclusion

Continued on the next page

Table 4 (continued)

No.	Title	Author(s)	Year	Decision	Reason / FE Code	IC1	IC2	IC3	IC4	IC5	Category
20	Iterative Network Pruning with Uncertainty Regularization for Lifelong Sentiment Classification	Geng et al.	2021	EXCLUDE	FE2: Lifelong sentiment classification on general review corpora, not medical or health text.	N	Y	Y	Y	Y	Formal exclusion
21	MASKER: Masked Keyword Regularization for Reliable Text Classification	Moon et al.	2021	EXCLUDE	FE2: General-domain reliable text classification without a medical primary dataset.	N	Y	Y	Y	Y	Formal exclusion
22	Transformer-Based Confocal Predictors for Paraphrase Detection	Giovannotti and Gamerman	2021	EXCLUDE	FE2: Paraphrase detection uses a non-medical primary task and dataset.	N	Y	Y	Y	Y	Formal exclusion

Statements and Declarations

Registration and protocol. This systematic review was not prospectively registered, and no formal review protocol was prepared. Therefore, protocol amendments were not applicable.

Competing interests. The authors declare no competing interests.

Funding. No funding was received to support this review.

Ethics approval. Not applicable. This study is a systematic review of published literature and did not involve human participants or animals.

Consent to participate. Not applicable.

Consent for publication. Not applicable.

Data availability. The structured data-extraction table, screening data, taxonomy assignments, and comparative analysis data generated for this review are available from the corresponding author on reasonable request. The study-level reasons for full-text exclusion are reported in Appendix Table 4.

Code availability. Not applicable. No analytic code was generated because the findings were synthesised narratively and no quantitative meta-analysis was conducted.

Author contributions. P. D. Manage: Conceptualization, Methodology, Investigation, Formal analysis, Data curation, Validation, Visualization, Writing – original draft. Y. Wu: Supervision, Validation, Writing – review & editing. J. Zhang: Supervision, Writing – review & editing. Y. Li: Supervision, Validation, Writing – review & editing.

References

- Alizadehsani R, Roshanzamir M, Hussain S, et al (2024) Handling of uncertainty in medical data using machine learning and probability theory techniques: A review of 30 years (1991–2020). *Annals of Operations Research* 339(3):1077–1118. <https://doi.org/10.1007/s10479-021-04006-2>
- Alnajjar M, Hamzeh R, Ghaith L, et al (2026) A hybrid NLP model combining biomedical entities and deep embeddings for automated identification of cancer types in biomedical texts. *Network Modeling Analysis in Health Informatics and Bioinformatics* 15:1. <https://doi.org/10.1007/s13721-025-00677-0>
- Andersen JS, Zukunft O (2022) Towards more reliable text classification on edge devices via a human-in-the-loop. In: *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART)*, Volume 2. SciTePress, Setúbal, Portugal, pp 636–646, <https://doi.org/10.5220/0010980600003116>
- Andersen JS, Zukunft O (2023) More sustainable text classification via uncertainty sampling and a human-in-the-loop. In: *Agents and Artificial Intelligence: 14th International Conference, ICAART 2022, Revised Selected Papers, Lecture Notes in*

- Computer Science, vol 13786. Springer Nature Switzerland, Cham, Switzerland, p 201–225, https://doi.org/10.1007/978-3-031-22953-4_9
- Baqraf Y, Keikhosrokiani P (2026) Enhancing trustworthiness of Arabic online health information quality evaluation using an enhanced BERT architecture with PCA and ICA feature weighting. *Scientific Reports* URL <https://www.nature.com/articles/s41598-026-43158-8>
- Beltagy I, Lo K, Cohan A (2019) SciBERT: A pretrained language model for scientific text. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp 3613–3618, URL <https://aclanthology.org/D19-1371/>
- Bressem KK, Adams LC, Gaudin RA, et al (2020) Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. *Bioinformatics* 36(21):5255–5261. <https://doi.org/10.1093/bioinformatics/btaa668>
- Campos MM, Farinhas A, Zerva C, et al (2024) Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics* 12:1228–1252. https://doi.org/10.1162/tacl_a_00715
- Can H, Yücedağ I (2026) User-level depression detection using long-context transformers and behavioral data. *IEEE Access* 14:73621–73645. <https://doi.org/10.1109/ACCESS.2026.3689467>
- Cardamone NC, Olsson M, Schmutte T, et al (2025) Classifying unstructured text in electronic health records for mental health prediction models: Large language model evaluation study. *JMIR Medical Informatics* 13:e65454. <https://doi.org/10.2196/65454>
- Chen X, Du Y (2025) Enhancing medical text classification with GAN-based data augmentation and multi-task learning in BERT. *Scientific Reports* 15(1):13854. <https://doi.org/10.1038/s41598-025-98281-9>
- Coban O, Yağanoğlu M, Bozkurt F (2024) Domain effect investigation for BERT models fine-tuned on different text categorization tasks. *Arabian Journal for Science and Engineering* 49:3685–3702. <https://doi.org/10.1007/s13369-023-08142-8>
- Desai S, Durrett G (2020) Calibration of pre-trained transformers. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp 295–302, <https://doi.org/10.18653/v1/2020.emnlp-main.21>
- Devlin J, Chang MW, Lee K, et al (2019) 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*, pp 4171–4186, URL <https://aclanthology.org/>

- Ding X, Lybarger K, Tauscher J, et al (2022) Improving classification of infrequent cognitive distortions: Domain-specific model vs. data augmentation. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop. Association for Computational Linguistics, Hybrid: Seattle, Washington + Online, pp 68–75, <https://doi.org/10.18653/v1/2022.naacl-srw.9>, URL <https://aclanthology.org/2022.naacl-srw.9/>
- Dong J, Jiang Z, Pan D, et al (2025) A survey on confidence calibration of deep learning-based classification models under class imbalance data. IEEE Transactions on Neural Networks and Learning Systems 36(9):15664–15684. <https://doi.org/10.1109/TNNLS.2025.3565159>
- Duan J, Jing H, Lu F (2025) Advancing disease prediction with data synthesis and noise-tolerant training. In: Proceedings of the 2025 IEEE 7th International Conference on Artificial Intelligence, Computer Science, and Information Processing (AICSIP), Hangzhou, China, pp 1–8, <https://doi.org/10.1109/AICSIP65423.2025.11427198>
- Durán-Silva N, Carrillo-de Albornoz J, Plaza L, et al (2024) BATRACIO: BASic, TRAnslational, Clinical, Research Phase Identification in BiOmedical Publications. IEEE Access 12:146476–146493. <https://doi.org/10.1109/ACCESS.2024.3463717>
- Ferdowsi S, Knafo J, Borisov N, et al (2023) Deep learning-based risk prediction for interventional clinical trials based on protocol design: A retrospective study. Patterns 4(3):100689. <https://doi.org/10.1016/j.patter.2023.100689>
- Ferrell B, Raskin SE, Zimmerman EB (2023) Calibrating a transformer-based model’s confidence on community-engaged research studies: Decision support evaluation study. JMIR Formative Research 7:e41516. <https://doi.org/10.2196/41516>
- Flores CA, Verschae R (2024) A hybrid method for clinical text classification based on confident predictions and regular expressions. In: Proceedings of the 2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC), Osaka, Japan, pp 64–69, <https://doi.org/10.1109/ICAIIIC60209.2024.10463358>
- Flores CA, Verschae R (2025) Active learning in biomedical text classification using a bag-of-regular-expressions approach. In: Proceedings of the 2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS), Madrid, Spain, pp 208–213, <https://doi.org/10.1109/CBMS65348.2025.00051>
- Flores CA, Figueroa RL, Pezoa JE (2021) Active learning for biomedical text classification based on automatically generated regular expressions. IEEE Access 9:38767–38777. <https://doi.org/10.1109/ACCESS.2021.3064000>

- Gal Y, Ghahramani Z (2016) Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning, pp 1050–1059, URL <https://proceedings.mlr.press/v48/gal16.html>
- Gao K (2024) Iterative multitask learning and inference from seismic images. *Geophysical Journal International* 236(1):565–592. <https://doi.org/10.1093/gji/ggad424>
- Gao S, Alawad M, Young MT, et al (2021) Limitations of transformers on clinical text classification. *IEEE Journal of Biomedical and Health Informatics* 25(9):3596–3607. <https://doi.org/10.1109/JBHI.2021.3062322>
- Geng B, Yang M, Yuan F, et al (2021) Iterative network pruning with uncertainty regularization for lifelong sentiment classification. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. Association for Computing Machinery, New York, NY, USA, pp 1229–1238, <https://doi.org/10.1145/3404835.3462902>
- Giovannotti P, Gammerman A (2021) Transformer-based conformal predictors for paraphrase detection. In: Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications, pp 243–265
- Greatrix T, Langbein FC, Whitaker RM, et al (2026) High-confidence labelling of pathology reports using LLM-based unanimous ensembles with limited data. In: Cafolla D, Rittman T, Ni H (eds) *Artificial Intelligence in Healthcare. AIIH 2025, Lecture Notes in Computer Science*, vol 16038. Springer, Cham, https://doi.org/10.1007/978-3-032-00652-3_27
- Gu Y, Tinn R, Cheng H, et al (2022) Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare* 3(1):1–23. <https://doi.org/10.1145/3458754>, article 2
- Guo C, Pleiss G, Sun Y, et al (2017) On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning, pp 1321–1330, URL <https://proceedings.mlr.press/v70/guo17a.html>
- Hendrycks D, Gimpel K (2017) A baseline for detecting misclassified and out-of-distribution examples. In: Proceedings of ICLR, URL <https://openreview.net/forum?id=Hkg4TI9xl>
- Hu Y, Khan L (2021) Uncertainty-aware reliable text classification. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. Association for Computing Machinery, New York, NY, USA, pp 628–636, <https://doi.org/10.1145/3447548.3467382>
- Huang K, Altosaar J, Ranganath R (2019) Clinicalbert: Modeling clinical notes and predicting hospital readmission. URL <https://arxiv.org/abs/1904.05342>

- Inan E (2026) Fine-tuning a cancer hallmarks classification model using large language model-synthesized resources. *Engineering Applications of Artificial Intelligence* 176:114702. <https://doi.org/10.1016/j.engappai.2026.114702>
- Islam MF, et al (2022) RNN variants vs transformer variants: Uncertainty in text classification with Monte Carlo dropout. In: *Proceedings of the 2022 25th International Conference on Computer and Information Technology (ICCIT)*, Cox's Bazar, Bangladesh, pp 7–12, <https://doi.org/10.1109/ICCIT57492.2022.10055922>
- Jang G, Kim MW, Nam YJ, et al (2026) EnergyRoute: Energy-based uncertainty routing for selective retrieval and large language model assistance in the hierarchical classification of biotechnology R&D projects. *Scientific Reports* 16:19697. <https://doi.org/10.1038/s41598-026-49886-1>
- Katsarou K, Jeney R, Stefanidis K (2023) MUTUAL: Multi-domain sentiment classification via uncertainty sampling. In: *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*. Association for Computing Machinery, New York, NY, USA, pp 331–339, <https://doi.org/10.1145/3555776.3577765>
- Khandokar I, Farghaly O, Kothari AN, et al (2024) Towards precision diagnosis: Integrating lexical analysis and deep learning for uncertainty detection and quantification in clinical reports. In: *Proceedings of the 2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS)*, Guadalajara, Mexico, pp 267–272, <https://doi.org/10.1109/CBMS61543.2024.00051>
- Kodati D, Dasari CM (2025) Detecting critical diseases associated with higher mortality in electronic health records using a hybrid attention-based transformer. *Engineering Applications of Artificial Intelligence* 139(Part B):109649. <https://doi.org/10.1016/j.engappai.2024.109649>
- Krishna NB, Lokesh K, Sekhar CS, et al (2025) A hybrid approach for predicting depression using deep learning. In: *2025 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*. IEEE, Piscataway, NJ, USA, pp 1–7
- Kumar A, Senapati A, Upadhyay R, et al (2025) An uncertainty-aware decision support system: Integrating text narratives and conformal prediction for trustworthy accident code classification. *Process Safety and Environmental Protection* 204:108134. <https://doi.org/10.1016/j.psep.2025.108134>
- Lakshitha B., Padmaja S., Suhana Tabassum, et al (2026) Sequential attention enhanced RoBERTa framework for multi-class emotional health classification. In: *Proceedings of the 2026 13th International Conference on Computing for Sustainable Global Development (INDIACom)*, pp 1–7, <https://doi.org/10.23919/INDIACom70271.2026.11526020>

- Lakshminarayanan B, Pritzel A, Blundell C (2017) Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems, URL <https://papers.nips.cc/paper/7219-simple-and-scalable-predictive-uncertainty-estimation-using-deep-ensembles>
- Laxminarayana J, Shaik MA (2026) Semantic-context fusion framework for robust sentiment understanding in clinical and medical text data. In: Proceedings of the 2026 9th International Conference on Inventive Computation Technologies (ICICT), pp 989–995, <https://doi.org/10.1109/ICICT68280.2026.11511119>
- Lechuga Lopez LJ, Elsharief S, Al Jorf D, et al (2025) Uncertainty quantification for machine learning in healthcare: A survey. In: Proceedings of the Sixth Conference on Health, Inference, and Learning, pp 862–907
- Lee J, Yoon W, Kim S, et al (2020) BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36(4):1234–1240. <https://doi.org/10.1093/bioinformatics/btz682>
- Lee S, Han YJ, Park HJ, et al (2025) Entity-enhanced BERT for medical specialty prediction based on clinical questionnaire data. *PLoS ONE* 20(1):e0317795. <https://doi.org/10.1371/journal.pone.0317795>
- Li Y, Liu C, Zhang Y, et al (2017) Enhancing binary classification by modeling uncertain boundary in three-way decisions. *IEEE Transactions on Knowledge and Data Engineering* 29(7):1438–1451. <https://doi.org/10.1109/TKDE.2017.2681671>
- Link NB, Huang S, Cai T, et al (2022) Binary acronym disambiguation in clinical notes. *International Journal of Medical Informatics* 162:104753. <https://doi.org/10.1016/j.ijmedinf.2022.104753>
- Liu X, Chen T, Da L, et al (2025) Uncertainty quantification and confidence calibration in large language models: A survey. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Volume 2. Association for Computing Machinery, New York, NY, USA, pp 6107–6117, <https://doi.org/10.1145/3711896.3736569>
- Liu Y, Liu Y, Mbuvi P, et al (2026) Mind the ambiguity: Aleatoric uncertainty quantification in LLMs for safe medical question answering. <https://doi.org/10.1145/3774904.3793027>, URL <https://dl.acm.org/doi/10.1145/3774904.3793027>
- Loftus TJ, Shickel B, Ruppert MM, et al (2022) Uncertainty-aware deep learning in healthcare: A scoping review. *PLOS Digital Health* 1(8):e0000085. <https://doi.org/10.1371/journal.pdig.0000085>
- Manage PD, Wu Y, Zhang J, et al (2026a) Overcoming BERT’s limitations in uncertainty: A novel two-stage solution for multi-class medical text classification. *Knowledge-Based Systems* 336:115315. <https://doi.org/10.1016/j.knosys.2026.115315>

- Manage PD, Wu Y, Zhang J, et al (2026b) BERT-TCPE: A term-class probability-enhanced BERT framework for uncertainty-aware medical text classification. In: Mei Y, Qian C, Bai Q, et al (eds) PRICAI 2025: Trends in Artificial Intelligence. Springer Nature Singapore, Singapore, p 531–539, https://doi.org/10.1007/978-981-95-7081-2_37
- Mena J, Pujol O, Vitrià J (2021) A survey on uncertainty estimation in deep learning classification systems from a bayesian perspective. *ACM Computing Surveys* 54(9):1–35. <https://doi.org/10.1145/3477140>
- Moon SJ, Mo S, Lee K, et al (2021) MASKER: Masked keyword regularization for reliable text classification. *Proceedings of the AAAI Conference on Artificial Intelligence* 35(15):13578–13586. <https://doi.org/10.1609/aaai.v35i15.17601>
- Moradi M, Samwald M (2021) Explaining black-box models for biomedical text classification. *IEEE Journal of Biomedical and Health Informatics* 25(8):3112–3120. <https://doi.org/10.1109/JBHI.2021.3056748>
- Naseem H, Catal C (2023) Domain specific transformers-based prioritization of re-admission for patients in healthcare. In: *Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval (NLPRI 2023)*. Association for Computing Machinery, New York, NY, USA, pp 51–56, <https://doi.org/10.1145/3639233.3639350>
- Naseem U, Dunn AG, Khushi M, et al (2022) Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT. *BMC Bioinformatics* 23:144. <https://doi.org/10.1186/s12859-022-04688-w>
- Nowak S, Biesner D, Layer YC, et al (2023) Transformer-based structuring of free-text radiology report databases. *European Radiology* 33:4228–4236. <https://doi.org/10.1007/s00330-023-09526-y>
- Page MJ, Moher D, Bossuyt PM, et al (2021) PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ* 372:n160. <https://doi.org/10.1136/bmj.n160>
- Peluso A, Danciu I, Yoon HJ, et al (2024) Deep learning uncertainty quantification for clinical text classification. *Journal of Biomedical Informatics* 149:104576. <https://doi.org/10.1016/j.jbi.2023.104576>
- Purcaru E, George M, Stammers M, et al (2026) From text to code: Leveraging machine learning for neurology outpatient clinical coding. *Neuroscience Informatics* 6(1):100257. <https://doi.org/10.1016/j.neuri.2026.100257>

- Qian W, He Y, Cai X, et al (2025) Granular-ball representation-based two-stage deep learning model for text classification. *Applied Intelligence* 55(18):1129. <https://doi.org/10.1007/s10489-025-07010-2>
- Reusens M, Stevens A, De Smedt J, et al (2024) Evaluating text classification: A benchmark study. *Expert Systems with Applications* 254:124302. <https://doi.org/10.1016/j.eswa.2024.124302>
- Rostam ZRK, Kertész G (2026) EQ-5D classification using biomedical entity-enriched pre-trained language models and multiple instance learning. In: *Proceedings of the 2026 IEEE 24th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, Stará Lesná, Slovakia, pp 391–396, <https://doi.org/10.1109/SAMI68106.2026.11420295>
- Rostam, Z. R. K. and Kertész, G (2026) Mutual learning and adaptive ensembling of pre-trained language models for hierarchical and flat scientific text classification. In: *2026 IEEE 13th International Joint Conference on Cybernetics and Computational Cybernetics, Cyber-Medical Systems (ICCC)*. IEEE, Piscataway, NJ, USA, pp 249–254, <https://doi.org/10.1109/ICCC68994.2026.11511643>
- Saqib HM, Mustafa G (2025) UMLS-augmented PubMedBERT for clinical note diagnosis classification: A comparative study. *Journal of Computing & Biomedical Informatics* 10(1). <https://doi.org/10.56979/1001/2025/1175>
- Sensoy M, Kaplan L, Kandemir M (2018) Evidential deep learning to quantify classification uncertainty. In: *Advances in Neural Information Processing Systems*, URL <https://proceedings.neurips.cc/paper/2018/hash/a981f2b708044d6fb4a71a1463242520-Abstract.html>
- Sinha A, Mickus T, Clausel M, et al (2024) Domain-specific or uncertainty-aware models: Does it really make a difference for biomedical text classification? In: *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics, Bangkok, Thailand, pp 202–211, <https://doi.org/10.18653/v1/2024.bionlp-1.16>
- Solarte-Pabón O, Montenegro O, Torrente M, et al (2022) Negation and uncertainty detection in clinical texts written in Spanish: A deep learning-based approach. *PeerJ Computer Science* 8:e913. <https://doi.org/10.7717/peerj-cs.913>
- Sun X, Lu W (2020) Understanding attention for text classification. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp 3418–3428, URL <https://aclanthology.org/2020.acl-main.312/>
- Šuster S, Baldwin T, Verspoor K (2023) Analysis of predictive performance and reliability of classifiers for quality assessment of medical evidence revealed important variation by medical area. *Journal of Clinical Epidemiology* 159:58–69. <https://doi.org/10.1016/j.jclinepi.2023.04.006>

- Tajrian M, Rahman A, Kabir MA, et al (2026) Leveraging Bayesian models to classify myths and facts in child development. *Discover Artificial Intelligence* 6:285. <https://doi.org/10.1007/s44163-025-00771-2>
- Tao X, Zhang M, Wang L, et al (2026) A multi-scale feature fusion module based on pre-training language models for medical text classification. *Pattern Analysis and Applications* 29:63. <https://doi.org/10.1007/s10044-026-01641-y>
- Ulmer D, Hardmeier C, Frelsen J (2023) Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. *Transactions on Machine Learning Research* URL <https://openreview.net/forum?id=xqS8k9E75c>
- Uslu EE, Sezer E, Guven ZA (2024) NLP-powered healthcare insights: A comparative analysis for multi-labeling classification with MIMIC-CXR dataset. *IEEE Access* 12:67314–67324. <https://doi.org/10.1109/ACCESS.2024.3400007>
- Wang Y, Beck D, Baldwin T, et al (2022) Uncertainty estimation and reduction of pre-trained models for text regression. *Transactions of the Association for Computational Linguistics* 10:680–696. https://doi.org/10.1162/tacl_a.00483
- Wang Z, Duan J, Yuan C, et al (2025) Word-sequence entropy: Towards uncertainty estimation in free-form medical question answering applications and beyond. *Engineering Applications of Artificial Intelligence* 139:109553. <https://doi.org/10.1016/j.engappai.2024.109553>
- Wang Z, Tian Y, Liu J, et al (2026) Hierarchical classification for differential diagnosis of fever of unknown origin: A multi-task learning approach with self-adaptive representation sharing. *Artificial Intelligence in Medicine* 171:103298. <https://doi.org/10.1016/j.artmed.2025.103298>
- Xiong M, Deng A, Koh PW, et al (2023) Proximity-informed calibration for deep neural networks. In: *Advances in Neural Information Processing Systems*
- Zabeen S, Hasan A, Islam MF, et al (2023) Robust fake review detection using uncertainty-aware LSTM and BERT. In: *2023 IEEE 15th International Conference on Computational Intelligence and Communication Networks (CICN)*. IEEE, Piscataway, NJ, USA, <https://doi.org/10.1109/CICN59264.2023.10402342>
- Zhan Z, Zhou S, Zhang R (2026) PEER: Towards reliable and efficient inference via patience-based early exiting with rejection. *Journal of Biomedical Informatics* 175:104988. <https://doi.org/10.1016/j.jbi.2026.104988>
- Zhang J, Yao W, Chen X, et al (2023a) Transferable post-hoc calibration on pretrained transformers in noisy text classification. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp 13940–13948, <https://doi.org/10.1609/aaai.v37i11.26632>

- Zhang N, Abbasi M, Khera S, et al (2026) Leveraging natural language processing and machine learning to identify chronic conditions from primary care electronic medical records. *Scientific Reports* 16(1):8441. <https://doi.org/10.1038/s41598-026-38594-5>
- Zhang Y, Li X, Liu Y, et al (2023b) A multilabel text classifier of cancer literature at the publication level: Methods study of medical text classification. *JMIR Medical Informatics* 11:e44892. <https://doi.org/10.2196/44892>
- Zhang Y, Zhao T, Miao D, et al (2025) Three-way multi-label classification: A review, a framework, and new challenges. *Applied Soft Computing* 171:112757. <https://doi.org/10.1016/j.asoc.2025.112757>